FIFRA SCIENTIFIC ADVISORY PANEL (SAP)

OPEN MEETING

FEDERAL INSECTICIDE, FUNGICIDE, AND RODENTICIDE ACT

DOCKET NUMBER: EPA-HQ-OPP-2017-0214

FIFRA SAP WEBSITE http://www.epa.gov/sap

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

CONFERENCE CENTER LOBBY LEVEL

ONE POTOMAC YARD (SOUTH BUILDING)

2777 SOUTH CRYSTAL DRIVE

ARLINGTON, VIRGINIA 22202

NOVEMBER 28 - 29, 2017

DR. TODD PETERSON: Good morning. I'm Todd Peterson and I will be serving as the Designated Federal Official to the US EPA Federal Insecticide, Fungicide, and Rodenticide Act, the Scientific Advisory Panel, which we commonly say as FIFRA SAP, for this meeting.

I want to thank Dr. McManaman, who is to my left here, for agreeing to serve as the Chair for the SAP for this meeting. I also want to thank both the members of the panel and of the public for attending this important meeting. We appreciate the time and effort of the committee members in preparing for this meeting, especially taking into account the holiday last week and your busy schedules.

In addition, I want to thank EPA's Office of Pesticide Programs and my colleagues on the FIFRA SAP staff for their hard work in preparing for this important review of EPA's continuing development of alternative high-throughput screens to determine endocrine bioactivity focusing on androgen receptor, steroidogenesis, and thyroid pathways.

By way of background, the FIFRA SAP is a Federal Advisory Committee that provides independent scientific peer review and advice to the Agency on

pesticides and pesticide related issues, regarding the impact of proposed regulatory actions on human health and the environment. The FIFRA SAP only provides advice and recommendations to EPA. Decision-making and implementation authority remains with the Agency.

The FIFRA SAP consists of seven members. The expertise of these members is augmented through the Food Quality Protection Act Science Review Board. The Science Review Board members serve as ad hoc temporary participants in FIFRA SAP activities, providing additional scientific expertise to assist in the reviews conducted by the panel.

As DFO for this meeting, I serve as the liaison between the FIFRA SAP and the Agency. I am also responsible for ensuring provisions of the Federal Advisory Committee Act are met.

The Federal Advisory Committee Act of 1972 established a system that governs the creation, operation, and termination of Executive Branch advisory committees. The FIFRA SAP meetings are subject to all of FACA's requirements. These include having open meetings, timely public notice of the meetings, and document availability. The documentation for this meeting is provided via the

Office of Pesticide Programs' public docket, which is accessible through the web at www.regulations.gov.

As the Designated Federal Official, for this meeting, it is a critical responsibility to work with the appropriate Agency officials to ensure that all appropriate ethics regulations are satisfied. In that capacity, panel members receive training on the provisions of the federal conflict of interest laws. In addition, each participant has filled out a standard government financial disclosure report. This is a confidential report.

I, along with our Deputy Ethics Officer for the Office of Science Coordination and Policy, and in consultation with the Office of General Counsel, have reviewed these reports to ensure that all ethics requirements are met. A sample copy of this form, for those who are interested, is available on the FIFRA SAP website. The address for that on the website is www.epa.gov/sap.

The FIFRA SAP will review challenging scientific issues over the next three days. We have a very full agenda and the meeting times on your agenda are approximate. Thus, we may not keep to exact times as noted, due to panel discussions and public

comments. We strive to ensure adequate time for Agency presentations, public comments, and panel deliberations.

For presenters, panel members, and the public commenters, please identify yourselves and speak into the microphones when it's your turn to speak. The meeting is being webcasted and transcribed and recorded.

Copies of all EPA presentation materials and written public comments are available at the public docket, again at regulations.gov. Copies of presentation material submitted this week by public commenters will be available in the docket within the next week.

For members of the public that have not pre-registered for public comments, please notify either myself or another member of the FIFRA SAP staff if you are interested in making a comment. At this time the agenda is full. However, as we move through the proceedings, if time allows, we may be able to accommodate additional brief comments for five minutes or less.

As I mentioned previously, there is a public docket for this meeting. All background

materials and questions posed to the panel by the Agency and other documents related to the meeting are available at the docket. Some documents are also available on the EPA SAP website. Access to those materials requires the docket number. This is noted on the meeting agenda.

For members of the press, EPA media relations staff are available to answer your questions about this meeting. If you want to be referred to a point of contact for that, please ask me at any time during the meeting.

At the conclusion of the meeting, the FIFRA SAP will prepare a report as a response to questions posed by the Agency, background materials, presentations, and public comments. The report serves as our meeting minutes. We anticipate the meeting minutes will be completed in approximately 90 days after the conclusion of the meeting.

Again, I wish to thank the panel for your participation. I'm looking forward to both a challenging and interesting discussion over the next three days. I'd like to turn the meeting over to the chair.

DR. JAMES MCMANAMAN: Good morning. I guess I can be heard. Welcome to the FIFRA SAP meeting, Continuing Development of Alternative High-Throughput Screens to Determine Endocrine Bioactivity Focusing on Androgen Receptor, Steroidogenesis, and Thyroid Pathways. Clearly the title reflects a large scope of possible interactions that we will be discussing today.

The way the meeting is organized is that there will be presentations by the Agency, followed by public commenters, and then we will turn over to the charge questions. During the Agency's presentation and the public commenters' presentation, the panel is free to ask questions of clarification. Following that, when we begin discussing the charge questions, it will be a discussion amongst panel members.

As Todd mentioned, this is being recorded, so if you forget to identify yourself as you begin to speak into the microphone, I will remind you, or maybe add your name myself, so that we can get everybody's information clearly identified.

I'm Jim McManaman. I forgot to introduce myself. I'm a professor at the University

of Colorado in Reproductive Sciences and Integrated Physiology. My expertise is generally in obesity, and I'm a permanent panel member. With that as my introduction, I'll turn it over to other panel members.

DR. DANA BARR: I'm Dana Barr from Emory University. My expertise is in exposure science of maternal child health.

DR. MARION EHRICH: I'm Marion Ehrich. I'm from Virginia Tech and I'm in pharmacology and toxicology.

DR. DAVID JETT: I'm Dave Jett. I'm from the National Institutes of Health and my expertise is in pesticide toxicology.

DR. SONYA SOBRIAN: Good morning. I'm Sonya Sobrian from the Howard University College of Medicine, Department of Pharmacology. My expertise is developmental neurotoxicology.

DR. SUSAN NAGEL: Susan Nagel, University of Missouri, Department of OB/GYN and Women's Health. My expertise is in endocrine disruption and steroid hormone action.

DR. THOMAS ZOELLER: I'm Tom Zoeller. I'm at the University of Massachusetts in Amherst. I

work on molecular mechanisms of thyroid hormone action and chemicals that can interfere with that.

DR. GRANT WELLER: I'm Grant Weller. I'm a senior scientist at Savvysherpa. It's a healthcare research and development firm in Minneapolis, Minnesota. I'm a statistician.

DR. KRISTI PULLEN FEDINICK: Good morning. I'm Kristi Pullen Fedinick. I'm a scientist at the Natural Resources Defense Council in our Health and Environment Program. My expertise is in population health biochemistry and the application of computational tools for risk assessment.

DR. EDWARD PERKINS: Hi, my name is Ed Perkins. I'm with the U.S. Army Corps of Engineers. My background is toxicogenomics, ecotoxicology, and in vitro screening for toxicology.

DR. REBECCA CLEWELL: I'm Rebecca Clewell from ScitoVation, which is a small research lab in North Carolina. My background is in development of in vitro and in silico tools to assist safety assessment.

DR. MICHAEL PENNELL: I'm Michael Pennell. I'm an Associate Professor of Biostatistics

at Ohio State University. My expertise is in statistical methods and tox risk assessment.

DR. IOANNIS ANDROULAKIS: I am Ioannis Andreoulakis from Biomedical Engineering at Rutgers University. My expertise is in systems biology of inflammation and endocrine hormones.

DR. SCOTT BELCHER: I'm Scott Belcher. I'm a professor at North Carolina State University. I'm with the Center for Health and the Human Environment. I'm primarily an expert in nuclear hormone action and endocrine disruptors.

DR. VERONICA BERROCAL: I'm Veronica Berrocal from the Department of Biostatistics, University of Michigan. My expertise is in statistical matters for environmental exposure and environmental epidemiology.

DR. J. DAVID FURLOW: I'm David Furlow, Professor of Neurobiology, Physiology and Behavior at the University of California, Davis. My expertise is in thyroid hormone and steroid hormone action molecular mechanisms, especially in development.

DR. JAMES MCMANAMAN: Thank you and welcome to all the panel members. With that, I think

we'll turn it over to the Agency for the first presentation.

DR. STANLEY BARONE: Thank you, Dr. McManaman and Chair. I'm Stan Barone. I'm the Acting Director of the Office of Science Coordination Policy. I'm a developmental neurotoxicologist by training. I'm here, basically, to represent the Agency and welcome you; and to underscore the importance of this peer review process and the standing panel and the ad hocs that are here today to join us in this robust dialogue to support the endocrine disruption screening program.

Your dialogue and the public comments today are going to be critically important as we move forward with improvements in the pivot to high-throughput testing and computational approaches in our screening program. I want to thank you all and I want to hopefully welcome you and hope you have an enjoyable time while you're here, deliberating here in Washington, D.C. Thanks.

DR. JAMES MCMANAMAN: Before we move on -- sorry. You know the old adage, out of sight, out of mind. We have a phone panelist, Dr. Shaw, on the

phone. Dr. Shaw, if you can hear me, would you go ahead and introduce yourself?

DR. JOSEPH SHAW: Thank you. I'm Joe Shaw. I'm at Indiana University School of Public and Environmental Affairs. My expertise is in molecular toxicology and toxicogenomics.

DR. JAMES MCMANAMAN: Thank you. Sorry for the omission.

DR. SEEMA SCHAPPELLE: Thank you. I also want to underscore the comments that Dr. Barone had just mentioned. I want to primarily -- can you all hear me?

DR. JAMES MCMANAMAN: Bring the microphone a little closer.

DR. SEEMA SCHAPPELLE: Sure. How is that? I want to welcome you all to the EPA. Thank you very much for being here. My name is Seema Schappelle. I'm with the EPA's Office of Chemical Safety and Pollution Prevention. And within the program here that I manage, we oversee the endocrine disruptor screening program. That's going to be our focus here for the next two and a half days. We're going to learn everything we want to about this program and more.

As you all know, EDSP is our program for assessing the potential for endocrine disruption when it comes to estrogen, androgen, or thyroid. We're doing this for pesticides, for chemicals, and environmental contaminants, not only in humans, but also in wildlife as well. That's very much the focus of our program and the crux of what we're doing here today.

I also want to thank all of you from the panel. I want to thank Chair McManaman for chairing this session, and all of you that have come here to offer your feedback throughout the next couple of days. It's very valued. Also, before I move on, I want to make sure to extend gratitude to the many scientists within the Agency and beyond that have extended their work and offered it to the program for utilization.

I'd like to start with our very own, with our scientists from the Office of Chemical Safety and Pollution Prevention and their work in assembling the models and applying them here within EDSP. I also want to thank our program office, EPA's Office of Research and Development, for the work that they've done, as well as members from NIH's NTP Interagency

Center for the Evaluation of Alternative Toxicological Methods, or NICEATM.

In all of the efforts that have been developed from our scientists, these are the tools and the high-throughput approaches and models that we are relying on and applying within the EDSP.

I just want to take a few minutes to give you just a little bit of orientation on what we do within the program, underscore the pivot that Dr. Barone had mentioned. In general, EPA intends to use the data that we've collected, either within the program or data that are available to us on pesticides and chemicals, to determine this risk for human health or the environment due to disruption of the endocrine system.

We're doing this in a couple of ways. Up on the screen you can see our approaches to incorporating computational toxicology data into the program. First and foremost, this is our ability to rapidly screen chemicals within the EDSP, within the universe of chemicals that we're looking at. I'll talk a bit about that as well.

We're also looking within the program to contribute to the weight of evidence screening

level determinations that help us determine a chemical's potential bioactivity. Also, this is a platform for incorporation of alternative data for specific endpoints within our Tier 1 battery, within EDSP.

It's important to remember that these are steps that we are taking to really try to achieve smarter testing of chemicals through the use of these CompTox tools and methods, not only for better prediction, but also the opportunity to reduce the reliance on animal testing and in vivo based toxicology studies that we know so well.

On the previous slide there was a blue circle representing the EDSP universe. Let me just break that down very briefly here. The suite of chemicals and substances that we're assessing here within the program consist mostly of pesticides, of actives and inerts that are shown here. Those are statutorily mandated for evaluation under the Federal Food, Drug, and Cosmetics Act. Then another portion of the substances that we're evaluating are mandated under the Safe Drinking Water Act.

You can see the breakdown on the screen of the number of substances that are contained in

those categories, and of course, kind of the unique number of substances that we're assessing here. That final breakdown is a little over 10,000 chemicals.

Lastly, I want to just pull on this slide that I think does a really nice job of not only showing our screening and testing approach -- which is depicted on the left side of this graph -- but also our incorporation of our pathway models and our use of alternatives and our intended use of alternatives moving forward.

On the top left of this graph you see the Tier 1 screening battery, the 11 assays that we utilize to determine the potential bioactivity of estrogen, androgen, and thyroid hormone systems. Using this screening battery on the top left under Tier 1. Based on the weight of evidence analysis that occurs on the Tier 1 battery results, substances that exhibit this potential for bioactivity with E, A, and T, then advance to the Tier 2 testing approach on the bottom left.

The Tier 2 tests are ultimately designed to identify any adverse endocrine-related effects that are caused by that substance, as well as

to establish this quantitative relationship between that dose and the endocrine effect.

That covers what's on the left-hand side. On the right-hand side, you'll see the ToxCast models that we've either proposed or that we will be talking about extensively here. These are really a depiction, on the right-hand side, of how our various pathway models are envisioned to work together. The ToxCast ER model, shown in red on the right, was announced by the Agency about two years ago in the summer of 2015. It's been proposed as an alternative to the three Tier 1 assays on the left, to the ER binding, the ERTA, and the uterotrophic assays.

As we move forward, we'll be assessing some of the additional approaches that we have proposed there. I have to apologize. My yellow stars, which I so cleverly placed, are misaligned. I'll talk you through that instead of relying on what's on the slide there. We do want to focus over the next couple of days on the three models which are shown there: the AR model, the steroidogenesis model, and thyroid, moving forward.

All of these are in various stages of development, as you'll see and as you probably have

seen from the white paper that has been submitted and released. For the purposes of our discussion here, the ToxCast AR model we are proposing as an alternative to one of the Tier 1 assays, to the AR binding. So, if you envision that first yellow star be on the fourth line down, AR model proposed. That's one of the things that we're going to be discussing extensively today.

With regard to the steroidogenesis pathway model, we'll be looking at a high-throughput H295R assay as an alternative to the low-throughput H295R steroidogenesis assay. Then we'll also be talking about and considering a broader pathway model. Again, we'll address that in depth today as well, and over the next few days.

And then regarding the thyroid. Within the context of EDSP, we've developed an initial framework in establishing our ability to utilize a network of AOPs for the potential evaluation of perturbation of thyroid function. These are all the pieces that we'll be talking about and I think this helps explain why and where they fit.

We're looking forward to the deliberations over the next two and a half days. I

again want to thank you all for being here and offering your feedback, your critical feedback as well. That's what we're here for. As we come out of this meeting, our intention is to take the feedback that we receive from you all, from our public commenters that are here today, and additional deliberations that will occur with our scientists within the Agency and beyond. Taking all of this information, bringing it together and utilizing it to make for a better program, to improve the approaches that we have and increase the robustness of what we are evaluating here within EDSP.

Thank you very much. With that, I'm going to turn it over to Dr. Bever who is going to start with our background on this paper.

DR. RONNIE JOE BEVER: Good morning. I'm going to present the background and basically provide you some context with --

DR. JAMES MCMANAMAN: Dr. Bever, could you move the microphone just a little closer?

DR. RONNIE JOE BEVER: Provide you with some context of why we're here, what we're doing, and what the Agency really expects out of this meeting.

I'd like to begin by telling you what I'll be describing today. I'll start with the problem. Endocrine disruptors represent a great concern and I will tell you how Congress has responded to this great concern by issuing mandates. These mandates resulted in the development of the Endocrine Disruptor Screening Program. Now this screening program discovered a great challenge in the amount of chemicals that we are actually required to test, and discovered that our rate of testing these chemicals was comparatively slow. I'll be describing that.

The endocrine disruptor screening program's response to that in the development of high-throughput methodology -- both for exposure and for bioactivity. Then I'll discuss some validation principles and the performance-based approach. Finally, I will summarize what the Agency would like to see with this meeting.

The problem: The Centers for Disease Control and Prevention estimated that 7.3 million women in the United States will seek the services of infertility clinics, based on data from 2011 to 2015. Women aren't the only ones suffering problems with fertility in this country. It is known that western

males in the western hemisphere are suffering from declines in sperm count; and, according to one meta-analysis published this year from the Human Reproduction Update, as many as 50 percent decline in the past 40 years.

Infertility isn't the only problem that endocrine disruptors can cause. And endocrine disruptors -- mainly the estrogens and the androgens -- will cause this infertility effect, but endocrine disruptors can also result in developmental problems. Once again, estrogens and androgens can play a part in that, but also thyroid hormones are particularly important with developmental problems.

These problems are notably of great concern; and in 1996, Congress issued the Food Quality Protection Act. This act amended the Food, Drug, and Cosmetic Act and the Safe Drinking Water Act. Now, it did a number of things, but we're here focusing on endocrine disruptors.

One excerpt from the Food, Drug, and Cosmetic Act is presented at the bottom of the screen. It calls for the Agency to develop a screening program using validated test systems, as well as other scientifically relevant information. We're basically

looking for the occurrences of estrogens, and also other endocrine-related effects which the Agency has specified to include androgens and the thyroid hormones.

The Endocrine Disruptor Screening and Testing Advisory Committee was formed in the same year, 1996, and generated this conceptual framework. I'm not going to go through all the bifurcations of this decision tree. Instead, I'm going to follow it down the way that the great majority of chemicals will go.

We start with sorting the chemicals and this depends on having information about the chemicals. Ideally, we will have bioactivity information, endocrine bioactivity information, as well as exposure information. This allows us to make judicious choices in our prioritization, which is the next step. It's very important, considering the number of chemicals we have, that we look at the most important chemicals based on bioactivity and exposure first.

After prioritization, we go to the Tier 1 screening. And the purpose of the Tier 1 screening is to identify endocrine bioactive compounds. Now,

this is a screening battery that includes in vitro and in vivo assays, as Dr. Schappelle showed in one of her slides. The decision on what is a bioactive compound is made by the expert opinion in the Agency based on a weight of analysis examination, weight of the evidence examination.

That means that just because there is some indication of endocrine bioactivity in a single assay, it's not necessary that the Agency will consider it a bioactive compound in need of Tier 2 testing.

Bioactive compounds are sent to Tier 2 testing. Tier 2 testing involves several in vivo studies -- which again, Dr. Schappelle showed -- and in Tier 2 testing, we decide which are endocrine disruptors and which are not. We also have established a dose response relationship at Tier 2.

At both Tier 2 and Tier 1, chemicals that show themselves not to be bioactive, or not to be endocrine disruptors, basically go to a holding bin. In this holding bin, there is no further analysis required at this time. Endocrine disruptors go to hazard assessment. After hazard assessment, of course, there is risk assessment. I just want to

point this out because we're considering exposure at the very beginning in prioritization, and we're considering exposure again during risk assessment.

Now, according to the EPA Science Advisory Board, the FIFRA Scientific Advisory Panel, the Endocrine Disruptor Screening and Testing Advisory Committee, and public comment, there was a universal agreement in the recommendation that a special program be formed in the Agency -- the Endocrine Disruptor Screening Program -- to address this important topic. I described the problem, so it's not only just as important as I'll show you, and as Dr. Schappelle has mentioned, we have a huge task; and so it's just apropos that a special program be developed to deal with it.

Now, once again, we use a two-tiered approach. After prioritization there is the Tier 1 battery, which serves to identify potential endocrine bioactive substances. That's followed by Tier 2 testing, which evaluates the dose response relationship and establishes if the substance is indeed an endocrine disruptor.

We have approximately 10,000 chemicals to deal with. Those include the inert and active

pesticides, as well as other substances to which humans will be substantially exposed. Dr. Schappelle actually did a breakdown of these 10,000 chemicals in one of her slides. The first list the Agency created to call in endocrine testing was 67 chemicals. The mandates allow the Agency to actually issue orders for test data. It also allows the Agency to stop the sale and distribution of chemicals should we choose.

Now, 1996 we have the mandate. In 2017, OPP has completed the weight of evidence analysis in Tier 1 for only 52 chemicals. At this pace it will take decades, millions of dollars, and sacrifice of a great deal of animals, to make it through all of these chemicals we're mandated to evaluate. Therefore, this Agency and this program sees it necessary that we develop high-throughput assays, a different way of looking at these chemicals, so that we can fulfill our mandate in a timely manner. These high-throughput assays will reduce cost, animal use, and testing time required, of course.

And they have the benefit of prioritization to be more efficient. That's important. Once again, prioritization is based off the data that we have for the chemical. So, if we

have no endocrine bioactivity data, it makes it more difficult to make a judicious selection, while high-throughput assays can and have supplied this bioactivity data.

Therefore, once again, as soon as the program can develop a high-throughput assay, working together with its partners such as ORD and NCCT, we will bring it to the attention of SAP. It may be that we will develop the model further and refine it later on. That's fine in the performance-based approach, which I'll be describing in a few minutes.

It's an easier task to develop the high-throughput alternatives for the Tier 1 in vitro test. The ER pathway model was developed, and already we accept it as an alternative for the estrogen receptor binding assay, the estrogen reception transcriptional activation assay. These are two in vitro assays. We also accept the ER pathway model as an alternative for the in vivo uterotrophic assay.

In today's meeting, we will be proposing an androgen receptor binding alternative, which is the AR pathway model. We'll also be proposing the high-throughput steroidogenesis assay as

an alternative for the low-throughput. We haven't made it to the aromatase assay yet, but we will.

Now I'd like to discuss, basically, our bringing these ideas of high-throughput exposure and bioactivity assays and models to the attention of the SAP, and our cooperation in working with the SAP through the years. I'll begin in January 2013. In this meeting we brought eight high-throughput estrogen receptor binding assays to the SAP. We also discussed using physical chemical properties to exclude substances from testing. This exclusion was based upon, for example, properties that would not allow it to be tested in the assays; for example, extremes in PKA.

Also, if the chemical would not trigger the molecular initiating event, it could be excluded. We could know that by, for instance, it's molecular structure. The estrogen receptor expert system quantitative structural activity relationship is basically an in silico model for predicting the chemicals that would successfully bind. The Agency proposed that these sort of concepts could also be used for androgen and thyroid evaluations.

The SAP's comments and what we did about them are shown here. The SAP suggested considering exposure information early. And once again, we plan on considering exposure during prioritization. High-throughput assays of bioactivity could benefit prioritization. Again, that's the study and that's what our plan is also. In fact, a few thousand chemicals have already undergone high-throughput assays for endocrine bioactivity. This sort of information, once again, allows us to make a judicious choice in which chemicals to test first, i.e. prioritization.

It was suggested that estrogen receptor assays needed refinement and additional assays for the suite. We addressed that when we came back in 2014, to the SAP. It was suggested that androgen receptor pathway model should focus on androgen receptor antagonism. The Agency agrees, and we will discuss that later on today.

They also said that the thyroid pathway will involve multiple modes of action. Of course, the Agency agrees, and we will illustrate in our framework several adverse outcome pathways.

In July 2014, we came to the SAP with high-throughput exposure models. We also discussed the high-throughput toxicokinetics and reverse toxicokinetics. ExpoCast is short for exposure forecast. It is basically a way of high-throughput quantitative estimates of exposure.

SEEM, the systematic empirical evaluation of models, integrates the predictions of multiple models and empirically evaluates model performance systematically over as many chemicals as possible.

The SEEM framework includes calibration and evaluation of the high-throughput exposure models using chemical concentrations found in blood and urine from the National Health and Nutrition Examination Study.

Now, high-throughput toxicokinetics predicts tissues concentrations based on oral dose, and reverse toxicokinetics they converse. This exposure modeling -- high-throughput exposure modeling, is important to us because, once again, without exposure there can be no risk.

The SAP comments for this was they were basically happy with the SEEM framework, but they

wanted the uncertainty reduced. They also generally concurred with our approach for high-throughput toxicokinetics and reverse toxicokinetics modeling. However, they suggested that we also look into the inhalation route, as well as the dermal route. Because once again, we concentrated initially on the oral dose.

In December 2014 we brought the ER pathway model. Now, this model consists of 18 orthogonal high-throughput estrogen receptor assays. We presented it as an alternative model for the estrogen receptor binding and the estrogen receptor transcriptional activation assays, and the uterotrophic assays. At this same meeting we presented our first generation AR pathway model. Once again, the AR pathway model has been brought to the SAP before, and we are presenting refinements to this model today based on recommendations from the panel.

Finally, we discussed the integrated bioactivity exposure ranking system as a measure of prioritization. Again, this is following our idea that prioritization should consider both bioactivity and exposure.

The SAP's comments on the estrogen receptor model was the following: First of all, they felt like it had several strengths, but they wanted uncertainty and sensitivity analyses done and they wanted more transparency. They wanted it to be more clear the methodology, as well as they wanted to be able to repeat the statistical analyses. Now, the Agency addressed all of their comments and we declared that we would accept the ER pathway model as an alternative for the ER uterotrophic assays.

In December, the SAP's comments for the androgen pathway model was -- and this is only a few of them. I'm going to be discussing more of their comments and more of our responses later. This is just a flavor of what is to come.

They said to evaluate cytotoxicity, and we addressed that. They said to expand the range of chemical structures tested in the assay battery. We've addressed that, and again, I'll detail this later.

They said, include methods to assess the potential effects of non-classical/non-genomic mechanisms that mimic or inhibit androgen bioactivity. This is a great idea, but the current low-throughput

androgen receptor binding assay does not do this. An alternative to that assay, therefore, would not have to do that either.

In general, the SAP was pleased with the integrated bioactivity exposure ranking system. The SAP suggested refinements to the IBER, including gaining a better understanding of how monitoring data would strengthen the approach; and secondly, increasing the NHANES exposure data integrated into the IBER model. Refinement of the IBER model is ongoing.

Now, the basic validation principles, however, have remained fairly constant. There was a meeting on validation in Solna, Sweden in 1996. They discussed topics such as this; and these same sort of topics, you will see in the EPA documents, and OECD documents, and other regulatory agencies.

First, there is relevance. OECD's Guidance Document 34 defines relevance of a test method as encompassing the regulatory need, usefulness of the alternative method, and associated limitations of the test method.

Fit for purpose is basically the context of use, as well as in the performance-based

approach, performance-based acceptance criteria, to determine if the model or the assay is actually meeting its purpose. I'll be discussing some of those criteria momentarily.

Reproducibility is defined in the GD-34 as the extent of -- reliability or reproducibility can be used as similar terms -- the extent of reproducibility of results from a test within and among laboratories over time when performed using the same standardized protocol.

In this instance, cross-lab validation is appropriate for any sort of method that you expect to be used in naïve laboratories all across the world. It needs to be such that each lab can get similar results. However, some of these 21st century techniques aren't conducive to being performed in naïve labs. These special techniques require specialized equipment and specialized training. Furthermore, when the performance-based approach is used, and the criteria are fully evaluated, the need for cross-lab validation may not be apparent.

Now, transparency means that we expect that the models are going to be -- our models, assays, the methodologies, will be readily understood by

everyone. We also want, as the SAP pointed out, that the statistical analysis be able to be reproduced by other people. It's important that we be very clear and comprehensive in reporting these things.

I'd like to talk about the performance-based approach. The performance-based approach is not a new thing. It was talked about in a paper in 1994, for instance. It's very different from what the EPA and most other regulatory agencies normally use. The regulatory agencies typically use a prescriptive method. They call it a prescriptive method because it gives you a very detailed methodology that you're expected to conform to in order to get the test results to be accepted.

Now, the performance-based approach handles it in a very different fashion. Instead of focusing on the methods, the performance-based approach focuses on the end result. That means it's very flexible in the methodology, as long as you stay within the same use context. The performance-based approach uses performance-based acceptance criteria, which I'll be discussing in the next slide.

It's very useful in that when you're using the performance-based approach, you're going to

be able to adopt scientific innovations immediately. How is this so? Because we expect these scientific innovations to give a better performance. In other words, they will meet the performance-based acceptance criteria, which means that they will be accepted in the regulatory context.

So performance-based acceptance criteria, it all comes down to having adequate reference chemicals so that you can really discern true criteria -- true measurements in the criteria. This means that systematic literature review is extremely important. Systematic literature review can identify a number of chemicals, and we need a number of chemicals. We need negatives and we need positives. Sometimes we need various types of positives, such as positives for agonist receptor binding, positives for antagonist receptor binding. With the high-throughput assays, you can look at many chemicals all at once, as opposed to the low-throughput; so that gives a very robust testing system.

Now, sensitivity is basically the proportions of positives correctly identified, while specificity is the proportion of negatives. Accuracy

is the proportion of correct outcomes predicted. Balanced accuracy is the average of sensitivity and specificity. These three factors -- sensitivity, specificity, accuracy -- are all examples of some acceptance criteria that could be used for fitness for purpose.

This is pretty important because if you look at cross-validation with the prescriptive method, you're basically telling the people how to perform the assay. Well, you want cross-lab validation to make sure that these labs performing the assay according to your methodology, they're generating similar results.

With performance-based acceptance criteria, if you're setting the criteria for accuracy at 95 percent -- just for the sake of argument -- you're making sure that these laboratories are giving you quality data in that regard, that they're giving you similar answers based on the wealth of reference chemicals that you supply.

Z-factor is also referred to as Z-prime. It examines signal dynamic range, as well as data variation. It can be a measure of reproducibility, as well as a measure of assay quality. It is different from what we'll be

discussing later, which is the Z-score. The Z-score is used a measure of cell stress. And again, we'll be discussing that later.

There are other criteria that can also be used for reproducibility. The most common of these might be percent coefficient of variation. These are just examples of some acceptance criteria that the performance-based approach could use.

Without cross-lab validation, it brings a huge benefit because cross-lab validation can take a long time to complete. It can take many animal lives to complete, and it's costly. So, if cross-lab validation is not necessary, it's good not to have it. This topic is not new either. NICEATM has already been discussing this and Warren Casey made a presentation about this.

I'd like to finalize my talk by telling you what our goals are. We are going to present a high-throughput androgen receptor pathway model, and we're hoping it's accepted as an alternative for the low-throughput androgen receptor binding assay.

We're also going to present a high-throughput H295R steroidogenesis assay, which is basically an upscaled version of the low-throughput

H295R assay, with some particulars that Dr. Paul Friedman will be discussing. We would like for it to be accepted as an alternative.

Finally, we would like the comments and advice from the SAP concerning our thyroid framework and our initial development of an approach to detect substances that can perturb thyroid function. We do hope to someday -- certainly not in this SAP -- but someday to develop some way of a high-throughput approach to detecting perturbations of thyroid function.

Thank you very much for your attention, and that's it for me.

DR. JAMES MCMANAMAN: Thank you, Dr. Bever.

Questions from the panel? Clarification questions for him before he leaves the seat?

DR. KRISTI PULLEN FEDINICK: I had one question about your statement -- and also in the white paper -- about not needing to incorporate the non-genomic mechanisms. Saying that if you have an existing test that doesn't test for specific

endpoints, the alternative then doesn't need to have that.

My question in thinking about it kind of metaphorically or an analogy, is if you have a bicycle as your initial test, but you built a car, is it appropriate to test that car on bicycle standards? Can you talk a little bit about why the Agency thought not about incorporating the non-genomic information into this particular replacement test for AR, for example? Or for other tests as you move forward?

DR. RONNIE JOE BEVER: Once again, we intend to bring forth what we believe as scientifically valid high-throughput systems as quickly as possible. We would like to use the performance-based approach so that any refinements that we make later can be incorporated and can be actually used. So when we're trying to establish an alternative -- I understand your approach, and that doesn't mean that we're not continuing to refine our products. We refined the estrogen receptor pathway model. We refine the exposure models on an ongoing basis. We're trying to perfect what we're doing.

But, at this stage, yes, I believe it's a valid point that the current binding assay, it

doesn't address this. So, if I'm just saying, hey, we're looking for it as an alternative for this assay; well, no, we don't have to.

DR. SEEMA SCHAPPELLE: Also, I'd like to just maybe remind the role of metabolism here, as well, in terms of further refinements as we go. There is more work to do there.

DR. THOMAS ZOELLER: What specifically do you mean by, orthogonal, when you talk about different assays?

DR. RONNIE JOE BEVER: I mean by orthogonal that they complement. And complement meaning that if there is a weakness in a particular assay -- for instance, let's say it only has a moderate sensitivity. There may be another assay that has an excellent sensitivity. Sometimes it's a caveat between sensitivity and specificity. But what we mean by orthogonal is that they complement each other.

DR. THOMAS ZOELLER: By complement I can think of a number of different ways two assays could complement each other. You just mean it in a general way, that it could be an assay for transcription versus binding, for example? Those would be complementary. Or it could be two different

transcriptional assays in two different cell lines, et cetera. So, it's a very broad --

DR. RONNIE JOE BEVER: I do mean it in a general way. Yes.

DR. STANLEY BARONE: Just to add to that in response to Dr. Zoeller's comments. We have different detection systems, we have different cell systems, difference species. So, within the battery, there can be lots of different types of orthogonal assays. That's also, as you pointed out, there are assays for different parts of the pathway.

One of the things that we've been exploring, along with our partners, is looking at the pathway in a more complete fashion. That's another aspect to the computational model. We're bringing in more of the biological construct data into the model, than just a simple, one single assay, or two simple assays, into the Tier 1 screening approach.

DR. JAMES MCMANAMAN: Other questions? Thank you very much.

DR. RICHARD JUDSON: Good morning, everyone. I'm Richard Judson. I'm with the Office of Research and Development, National Center for Computational Toxicology, and I'm a bioinformatician.

I'm going to tell you about what I think is some complicated biologies, but some very simple math. My goal is to explain it so that all of you -- and you're a very diverse group of folks -- sort of understand all of it without insulting the intelligence of the experts in your respective fields.

My outline here. I'm going to state what our objective is. We're not trying to solve all problems. We have a specific objective. We had to develop our own set of reference chemicals for the androgen receptor. Then I'm going to talk about the technology of the model itself and the assays behind it, some results where we go through this validation process, and then a little discussion.

Before I do that, I want to make a comment on Dr. Pullen's question about the non-genomic mechanisms, which we don't address here. After the last SAP, talking about the estrogen receptor model -- which is perfectly relevant here -- the question of, why don't you see if you can't detect these?

We did a little bit of searching for enough chemicals -- you can't validate an assay, or test an assay, or test a model without some reference chemicals. We didn't spend a lot of time, but we

never found a set of sufficient number of reference chemicals to feel like we could do that test. So, if there are people out there who say, I can give you 10 or 20 chemicals that act through non-genomic mechanisms for ER or AR, we'd love to hear about that.

Having said that, let's move on. Our objective -- this was six or seven years ago when we started this process -- was not to replace assays, but simply to prioritize chemicals. We have this list of 10,000 chemicals which we could not run through the standard Tier 1 battery. So, can you simply do some high-throughput screening to do prioritization? That's our first goal.

You don't need to know the exact truth, whatever the truth is, to do a pretty good job of prioritization. The method has to be able to test thousands, up to tens of thousands of chemicals, which I think we're doing a pretty good job at. But then as we showed that we had pretty good accuracy relative to the Tier 1 test. We did this for ER and now we're doing it for androgen. We said, okay, are we good enough that we can use the high-throughput method as an alternative for the -- not a replacement, just an

alternative for the Tier 1 test. So, there's two goals here.

The overall approach is we're going to integrate multiple high-throughput screening assays. I'm going to spend a lot of time talking about one assay -- there is no one perfect assay. You have to use multiple assays. We can talk more about what orthogonal means, if people want.

The reason you have to have more than one assay, and why this orthogonal idea is important, is because chemicals can -- we call interfere. They can cause false-positive activity that has to do with the assay technology and nothing to do with -- in this case -- with the androgen receptor activity, at all.

Different chemicals can cause different kinds of interference in different kinds of cells and readout technologies and so on. We're going to apply this performance-based validation. We're going to take a set of reference chemicals that we define upfront, and where we sort of know the truth about active/inactive, weak/strong, and see how well this integrated model works against those.

The reference chemical effort -- and this was really driven by Dr. Kleinstreuer at NTP --

but she led an effort to go through a systematic literature review of in vitro data from the literature where you want -- essentially what you want is chemicals that have been tested in many labs and get consistent results. Either they're positive as an agonist or an antagonist, or they're always negative. So you're always bootstrapping your way looking at the history based on the literature.

In order to do this, the systematic part of this says that you extract really detailed information about the studies to make sure the studies were well done. I'll talk a little bit about that. We also brought in chemicals that had been validated in some way by ICCVAM and ECVAM and OECD and other agencies.

The literature effort -- I don't think it's on here, but I think every paper had to be read by two people. It was a real lengthy, time-consuming effort. Each paper, okay, what is the PubMed ID, the author, the year, make sure that you know what chemical is being tested. All of you who deal with chemicals know that it's easy to -- people call chemicals lots of different things, so they had to

guarantee that the chemical tested in this paper and this paper were really the same chemical.

Then you want to know do they call it active/inactive, any sort of notes about the response, and then a quantitative value, which could be an AC50, IC50, relative binding affinity and so on. What the assay is, what cell type, what the cell culture media was, what the readout type was. Is it a fluorescence based assay, is it a radioligand binding assay, and so on. What reference controls.

You always sort of validate your assay in your lab by running a positive control and a negative control. How many doses and so on. Especially for the antagonist assays, what is the cytotoxicity? And for those of you who don't know about antagonist assays, essentially you take an agonist assay which starts low, you add the chemical, it goes up. An antagonist assay, you put a reference agonist in, you start high, and then you put your antagonist in, which will then displace the agonist and you go low. Since going high to low, you can also have that happen because of cytotoxicity. Cytotoxicity can mimic antagonistic activities. We have to control for that.

The result of that long process was they identified 103 chemicals. This is for the binding data here, 1,100 rows. A row is a report of one chemical and one paper. Obviously, on average you found about 10 reports of chemicals, but there are certain chemicals where you find 40 or 50 or 60, and some where you only find one or two. Transactivation assays, there are 135 chemicals to choose from. This was the body of data, and then there was some selection criteria that brought this down to a final set that was used.

This is just an illustration -- it's going to be a little hard to see -- but what the raw data looks like. This is all of the 103 binding assays or some subset of those, so each column is a chemical. The dots correspond to the -- if it's a colored dot, what is the potency, so the AC50 or the IC50 of that chemical. And an important point to note. Note the scale, the distance between the major tick marks is four orders of magnitude. You can see the positive results can span one, two, three orders of magnitude. So there is a lot of variability in the literature data, which that's just inevitable because

different assays have different sensitivities and so on.

So, you're starting off with some uncertainty about what the true potency is, but then the black dots – which, with my old eyes, it's hard to see black versus the other colored ones -- those would be the upper limit of testing. I tested to 100 micromolar, and I did not see any activity. I call it negative.

You can see even some of these positive chemicals that have multiple positive reports, have reports that are negative. My lab tested higher than your lab and I called it negative. There is some uncertainty about, is this chemical even active against this target. You start off with the literature has this uncertainty, which is – that's just the truth.

Once you apply the acceptance criteria -- which I'll talk about in a slide or so -- these are the final chemicals that we use. It's a smaller set of chemicals and notice the scale is quite a bit compressed. The major axis difference is just one order of magnitude. But for most of these chemicals, most of the activity spans one to two orders of

magnitude. But still, there is that uncertainty in what the true potency is.

Just in case you're wondering, yes, there are big dots and little dots. The big dots, there were a bunch of reports from the literature had roughly that same concentration.

To be accepted as an agonist reference chemical, you had to have at least three experiments and at least 70 percent yield positive results. What that means is that if you only had three reports, all three of them had to be positive. If you had four reports, one of them could be negative. That would be 75 percent would be positive. Then we put them into these bins of strong, moderate, and weak. It gives you some kind of qualitative evidence of what the potency is.

The negatives had to have at least three reports that were negative, and there could not have been any positive reports. Later on you'll see some chemicals labeled as moderate/weak or moderate/strong, and that just means again, there is some uncertainty in what the potency is for the chemical based on the literature data, and so the uncertainty can span these groups.

The antagonist is roughly the same. You had to have greater than three experiments, 70 percent yield positives. A lot of chemicals are really weak, so the antagonists in the literature tend not to be really blazing potent. They're not like the natural hormones. They're not like testosterone.

We had a mixed batch. There was not a lot of literature negatives, so we had to fall back and only have two or more negatives to call something a negative, to have a sufficient number.

Having used that criteria, there are finally a set of 54 chemicals; 37 agonists, 28 antagonists. There is a number that overlap because they were consistently positive as antagonists and negative as agonists, or vice versa. Of those 54, 46 had overlapping data with the model. So it's that 46 that we're finally going to use as the validation process. If you want to look at the chemicals, they're in Table 2.2.

That's the reference chemicals. Any questions about that, clarifying, before I go on? Great.

Now I'll jump into the model. I already mentioned this, the issue, the reason for the

model is because no assay is perfect. There are three kinds of imperfection. One is that each of these different assays test some slightly different piece of the biology. We're looking at different places along the pathway, but also, we have different cell types. You have differential sensitivity from one cell type to the next, or are you a 96, 384, or 1536 well plates and so on. So, you have different sensitivity, and then how we actually measure the signaling is different.

Then we have assay interference. Just as an example that always makes sense to me, is if you have a radioligand binding assay, which we have a number of. So, there is cell-free, you just have the receptor sitting there in a solution and it has a radio label, testosterone or some other ligand in there. And you put in a potent other androgen, it will displace some of that. You have some radioactivity in the solution, and your potency or your efficacy is a measure of how much radioactivity is there.

Now you throw in a chemical which denatures the protein, so the protein simply dissolves, and all of the radioactive stuff jumps out

in the solution and you actually get a very nice binding curve, but it has nothing to do with binding to the receptors. That's the sort of thing we have interference.

Each of the different technologies has its own separate kind of things that could interfere with it. Both of those are -- if I run the assay today, or I run it tomorrow, or I run it in my lab and you run it in your lab, we will get the same false-positive results.

Then there is also noise. I run it today and I get one answer, and I run it tomorrow and I get another answer. All of these assays, just the high-throughput way, we run them once. There is some noise here.

What we're going to do is have a set of different assays across different points in the pathway, and we're going to use a relatively simple mathematical model to integrate all of that. What we finally come up with is a composite dose response curve for agonism, antagonism, and these different interference modes. And I'll talk more about that.

I just have a couple of slides on the experiments. We had 1,855 chemicals that we purchased

-- the EPA purchased -- through a vendor, Evotec. They were put on plates and shipped to all the vendors, so everybody is actually working with the same samples. There is a chance that actually there is a wrong chemical at the vendor, they mixed something up, so we'll get wrong answers across the board; but we see consistency and there may be some of that here.

Chemical QC, is the chemical in the well what we wanted to be in the well. That's an ongoing process. It turns out, you take 1,800 chemicals and ask any vendor to say, okay, what's actually there? That's a really hard problem. We're about 60 percent of the way through getting that done. We're continuing to pull that in. Again, there could be some false calls here because of that, but the bottom line of the QC is that most of the chemicals are what they're supposed to be and they're about the right concentration, what they're supposed to be.

The chemicals run in 11 assays, which the next slide shows. Out of each assay we get a -- hitcall, was it positive, negative, and then what was the potency. The AC50, it's the concentration at half-maximum activity. Everything except the

NovaScreen, which are the cell-free binding assays, were run in concentration response. With NovaScreen we did a single screen at relatively high concentration, 40 micromolar. If it was relatively potent, we would go run it in concentration response.

The reason for this is purely cost. These are really expensive assays and they were run in the context of screening. We had hundreds of assays NovaScreen ran because this is part of a general looking at all kinds of pathways. All the data is passed through our in-house open source ToxCast data pipeline, which we and lots of other people are using.

Don't look at this, but if we have questions later on about what the assays are, it's in your slide packet and we can go back. The major points, the different assays are looking at different points of the pathway. They have different cell types, some are the full-length receptors, some are the ligand binding domain.

The first assay, the human androgen receptor cell-free assay, is the only one that is a mutant protein. And we've been criticized about using this. It's a well-known mutant which is a little more

open pocket, it's a little more promiscuous, so it tends to have more positives than others have.

There is no technical reason for doing that. It has to do with a patent dispute. The company, NovaScreen, couldn't get a license to the patent for the wild type human receptor. There are lots of technical and economic reasons for some or all of this.

The key points of the model; so, somebody can say, you've run 11 assays in a chemical and if any one of those assays are positive, this is an androgen and we should act on it. So, that might be the real conservative approach. But the fact is, no, that's not right. As I've talked about, there are a lot of ways to get false positives, and so we need to look across multiple assays and get some kind of a weight of evidence.

What the goal of the model is doing is not to give you a final potency estimate, but it's really to try to distinguish false from true activity. What we're going to do is this simple mathematical model. We're going to classify a chemical as a true agonist, a true antagonist, or it's acting through one of several defined interference modes. We quantify

each of these modes -- agonist, antagonist, and so on -- with this area under the curve value -- I'll have a little picture later on -- and the mode with the highest AUC is selected. The AUC is not potency, but we give you potency estimates. Again, more on that later.

For those of you who are nuclear receptor biologists, just don't laugh. This is my naïve understanding of how nuclear receptor biology works. The way this works is you have the actual receptor. A chemical actually binds to the receptor itself. Then it dimerizes -- two dimers come in. We don't have any dimerization assays, but in principle you could develop one. The dimer recruits some cofactors and forms the mature transcription factor. That then goes and binds to the DNA, and of course, DNA then you create RNA and then the RNA turns into protein. That's the standard cycling, or the biology of the nuclear receptors. These colored circles are kind of the underlying biology. That's agonism.

Antagonism, you actually bind, dimerize, recruit the cofactors, and you bind to the DNA, but you bind in a way that halts -- doesn't allow

transcription to happen. You actually can measure that.

For each of these modes, the underlying biology, we have one or more assays. These white stars are the actual assays and the lines connecting them, the arrows, indicate that if you actually bind, you should light up assays one, two, and three. If you recruit cofactors, you should light up assays four and five and so on.

So that's the true biology. Then we have these interference nodes. I called them pseudo-receptors initially and people hated that, so we just call them nodes now. But there is some process that if you activate that process, you cause the assays for that particular technology to light up. Really, all the model is doing is looking at different patterns of which assays light up and say, can I explain that pattern, either by true agonism, true antagonism, or one of these interference modes?

Essentially, we've laid out 13 or 14 -- I don't remember the exact number -- of alternative hypotheses. The model tests each of those hypotheses and says, which one has the greatest evidence, and we assign that.

If you imagine we have a true agonist, it goes to the true agonist and it lights up all of those agonist assays. If we have a true antagonist, it lights up the top set and then the bottom left. If you have interference up here, it will only light up those two, so you will see that pattern of activity.

The model, this is the whole model. That's all there is to it, but let's go back here for a second and say, all right, if I have a true binder and it really binds here, I can now predict what these assays are going to do. I bind here, these assays turn on. That's sort of the forward model, or the true biology. But what we know is actually the assays and we need to guess what the receptor is.

The first equation it says, if I know what the receptors are, I can predict what the assay results will be. And the R here is simply how high up the activity curve you are for that assay. F is one if there is a black arrow connecting the assay and the receptor; and it's zero otherwise. We want to invert this thing and take the assays and guess the receptor.

The easiest way to do that is, essentially, the computer guesses the value for the receptors. It guesses, it's a little bit of this and

a little bit of that. And then it predicts what the assay results should be. Then what you do is you calculate the error, the difference between the predicted assays and what we actually measured. So, you just iterate over and over again, in a smart, gradient, driven way, and you minimize this error. You can finally drive it down to very low levels.

The error term -- so, we have a gradient minimization method that minimizes this error. It has the basic squared error, but then it has this penalty term. A technical issue here is this is what's called an underdetermined system.

We have many more variables, many more of these R's than we have assays, A. There are actually many solutions to the equations. They give the same minimum error. So, in a sense, we're free to choose those. And what we do is we say, we prefer solutions that have the chemical doing one or two things, rather than six or seven.

You might have a solution that has many of these receptors lit up, and it gives you a low error, or one that only has one of the nodes lit up. All this penalty term does is it drives a preference to the most parsimonious solution.

Finally, once you have that -- and you do that at every single concentration and then you knit a curve together, and you calculate this area under the curve -- which is just the area under the curve, but it has this funny term -- I've lost my -- it sort of wanders around. Anyway. It's a sign of the slope. I'll show you that graphically as I go along. It's easier to describe it with a picture. That's all the model is.

Just another thinking, in a flow-charty way -- we do this separately for each chemical, and for each chemical we do it for each concentration. We take the measured assay values, the efficacy, how high up the curve you are, and we guess the initial value for the receptors and we just iterate until you minimize this error. Then you finally draw the concentration response curve.

Here, this is an example of the actual assays. We see the assays rise up nicely -- and this is Bisphenol A in the actual antagonist mode. Then the composited antagonist curve is this red curve. That's what the model finally says, is that's the probability, if you like, of antagonism.

We have this other curve, the black one, which is one of the other modes. It's actually the binding -- the cell-free binding. It's the one up at the top of the pathway and it's just because, notice the black assays turn on a little earlier than some of the others.

The model says, okay, early on at low concentrations, maybe it's an antagonist, maybe it's this binding interference; but if you get up more and more, higher in concentration, the other assays turn on, the probability of this goes down. So the area under the curve, this sort of virtual area under the curve, you have a positive integral. Then as it turns down, it's a negative integral. In the end, this curve, this black curve, has an area under the curve of close to zero. Whereas, the red is relatively high.

Just a little more on what these curves look like. The antagonist curve is red, the agonist is blue. I'll show an example later on. We have lots of cytotoxicity assays. There are 33, I think, that are used here. We define a region where cytotoxicity is going on. We'll use that to qualify the results later on. That's this gray region. The red bar is

the center of that, where most of the cytotoxicity is happening.

The green bar is what we call the pseudo-AC50. The simplest models, you just take the assays and you just add them all together and say, where is the average log AC50? That's what this green bar is. You can see it can be -- it's not going to necessarily be at the same point halfway up the antagonist curve.

Finally, the area of the curve. It really is just the integral under this -- there is an integral under the red curve, there is an integral under the black curve. There is no magic about that. It's just calculate how much is under there. It's roughly going to be proportional to the log of the AC50.

The model scoring -- so, now we have these AC50s. We calculate one for the agonist, one for the antagonist, one for all of the interference modes. Roughly say, if the AUC for agonist or antagonist is greater than 0.1, we're going to call it positive for that mode. If it's less than 0.001, it's definitely negative. If it's in the middle, it's inconclusive. We can't finally say yes or no for all

of these chemicals. There are a bunch in that middle region.

Again, there is nothing magic about this 0.1. You can make kind of a theoretical argument that 0.1 is about -- corresponds to an AC50. If all of the assays were positive with an AC50 of 100 micromolar, the AUC would be about 0.1. Since we test our assays up to 100 micromolar, if you're above 0.1, we have no idea what's going on there. We just don't have any evidence.

I've mentioned that antagonist cytotoxicity is a confounder, so we do two kinds of cytotoxicity filtering. The first is the antagonist assays, which are both run at NCGC up in Rockville. It's an NIH lab. They ran a concurrent cytotoxicity assay. It's essentially run the same time, the same day.

If the AC50 for the assay -- the antagonist assay -- was greater than for the cytotoxicity assay -- so, cytotoxicity turns on first and then the antagonist, we would just call it inactive and set the AC50 to a million and the top to zero. We filter those out. That's a hard cutoff, but those assays, like every other assay, are not perfect.

We find lots of cases where cytotoxicity assays can have false positives and false negatives also.

Then we do general cytotoxicity filtering. This was all explained in a paper from our group two or three years ago. We showed that looking across 1,000 chemicals with hundreds of assays, that all chemicals, if you get up to the cytotoxicity limit, you have this big burst of activity. Lots of assays turn on. Probably having nothing to do with the receptor they're trying to measure. We have a hypothesis about why that is. What we do is we calculate this z-score which I'll illustrate.

A low z-score essentially means the activity is happening in the region where cytotoxicity is happening, so you should proceed with caution. A high z-score means that at least cytotoxicity can't be the explanation for false activity. So, we'll put that into what we call the confidence score.

This is just an illustration of this. Here we have three example chemicals, and every chemical, if you just plot a histogram of where the hits -- where the AC50s are, just plot that, you see this burst of activity. We saw that as just a phenomenon that was very common. Then we realized

that that burst of activity is happening at the concentration where cytotoxicity is happening. We were able to actually just take the cytotoxicity assay and predict where you would have this burst of activity going on.

We certainly had many examples of chemicals that lit up; estrogen, androgen, and other assays that couldn't be estrogens or androgens. We think it's really because of this cytotoxicity false-positive effect. We see chemicals without the burst and without cytotoxicity, and we just presume that that's occurring at concentrations above 100 micromolar where we don't test.

This is the concentration micromolar scale. What we do is we simply shift all of these curves so that the center of cytotoxicity is zero, and we call that the z-scale. This is actually not correct. The zero should be here on the scale instead of the minus three. But then the gray area, the cytotoxicity area goes over, essentially, three standard deviation, or three median absolute deviation away from the center.

So, just a kind of hand-waving explanation of why you get this false activity in, for

instance, the androgen receptor assay. Imagine you have a transcription assay, and so you have the actual receptor, you put a chemical in, the chemical binds to the receptor; and what this does is it, again, it forms the transcription factor, it binds to the DNA, it causes activation of a particular gene. These cells have been engineered so if that gene is activated, you get a red glow or a green glow. That's how these assays work.

Now you have some chemical which does not bind to the assay. You put a lot of it in there, enough that it makes the cells really sick. As cells get sick, they regulate all sorts of activity trying to -- maybe they're going through apoptosis, maybe they're trying to respond to that stress and trying to recover from that and so on. All sorts of genes are turned on that are not necessarily turned on because of the particular transcription factor that they were supposed to be activated with.

Accidentally, you turn on that gene that has the reporter attached to it, and so you get this sort of accidental activity that has nothing to do with binding. We see chemicals that in that region are causing oxidative stress, they're reacting with

DNA, they're reacting with proteins, they're goofing up the mitochondria and so on. Lots of things are going on in that cell stress region.

How we do the cytotoxicity, we actually have 33 assays with multiple cell lines and primary cells, and we look at direct cytotoxicity as well as -- we have proliferating cells and it's the rate of proliferation just coming down. Chemicals deemed cytotoxic -- if two or more assays were active, then we calculate the cytotoxicity median, which is the median of the log, which is the red band. And then we have the MAD, median absolute deviation. I can go into that in more detail than you care for. This gray region is 3 MAD from the median. That's how that is done. I think I'm almost done with the methods finally.

Okay. So, confidence scoring. We calculate the AUCs, which is very automated and there is just a formula that does that. But then we have these other factors that are especially important for the antagonist. We have the concurrent cytotoxicity I already talked about. We have this general cell stress, the z-score. We have the antagonist confirmation data, which is this next point.

So, what we did is, we took one of the antagonist assays and we ran with two different concentrations of the reference agonist. Again, the antagonist assay, you just take an agonist assay, you turn it on with an agonist, and then you put an antagonist in which displaces the agonist and brings the response down.

If you have more of that reference agonist, it takes more of the antagonist to have an effect. You get right shifted. You shift your AC50s to higher concentration if you have higher responses. A true antagonist, you will see the shift in its AC50 between the low and the high -- in this case, R1881 concentration. Whereas if the activity is being caused by cytotoxicity, the AC50 should be exactly the same. It has nothing to do with that. Or by some other kind of assay interference. You want to have the shift in the right direction for a true antagonist.

Finally, the confidence score is if you have the -- and this is all for the antagonist. The AUC for antagonism, R2 is greater than 0.1, you get two points. If it's in that ambiguous region, you get a point. If the average z-score is greater than three

-- which means on average, all of the assays are turning on outside of the cytotoxicity region -- you get a point. If you get the true shift, you get three points. That's really confirmatory that you're going in the right direction. You see the good dose response curves for both concentrations.

You can actually -- if a chemical is pretty weak, the second concentration of R1881 may have shifted it so high that we don't actually see that. If you go from a hit that's pretty weak, to a no hit as you add R1881, we give it two points. It's still probably going in the right direction.

If you get a shift that's in the right direction, but again, there is uncertainty in the actual potency of the chemical, we still give it a point if it's going in the right direction. But you can't definitely say that one is greater than the other. Finally, if it's going in the wrong direction, we take a point away. A chemical can have a score from negative one to six.

Any questions on lots of detail on the methods? Any questions there before I go on? Yes, ma'am?

DR. KRISTI PULLEN FEDINICK: I had a question about your NovaScreen. Did you run those samples -- the 40 micromolar samples -- did you have replicates of that, or is it a single run at 40 micromolar and then --

DR. RICHARD JUDSON: It's run in duplicate.

DR. KRISTI PULLEN FEDINICK: In duplicate, okay. Do you ever sample from the chemicals that didn't pass that screen? Looking at, say, the non-concentration response chemicals, and then go back just to make sure that that 40 micromolar is correct?

DR. RICHARD JUDSON: We did that, not necessarily for this assay. Again, we tested a thousand chemicals by 300 assays there. We took a random collection of negatives. All of the assays were run in this 40 micromolar. We missed something. I don't know whether it's a few percent, so we could have missed things, yes.

DR. KRISTI PULLEN FEDINICK: One final question. Did the NovaScreen act as a screen for the other assays? So, you just ran those in the single concentration and then -- okay.

DR. RICHARD JUDSON: Right, so every other assay was -- all the assays were run totally independently.

DR. KRISTI PULLEN FEDINICK: Great, thank you.

DR. JAMES MCMANAMAN: Other questions? Marion?

DR. MARION EHRICH: A basic pharmacology question. It looks like everything here was strict competitive antagonism?

DR. RICHARD JUDSON: Yes, I think that's true, right.

DR. MARION EHRICH: There can be other kinds sometimes.

DR. RICHARD JUDSON: We would welcome suggestions on what else to look for.

DR. IOANNIS ANDROULAKIS: Just a clarification regarding the first example that you showed. Is there any way -- or how do you differentiate between a non-specific massive event like what you described, and maybe an indirect or a more complex mechanism that might lead to an activation of a more non-specific event that will actually have a measurable end result?

DR. RICHARD JUDSON: Do you know what? I spent probably a year or 18 months trying to solve that problem. The assays are all of these different technologies and they're happening at slightly different concentrations. And so saying, what happens first and what happens second -- which is part of that -- the data is just too noisy to really sort that out. It's an interesting question.

DR. JAMES MCMANAMAN: In your assays, I think you start off by taking those compounds that you feel have the structure to be an agonist or an antagonist. Is that incorrect?

DR. RICHARD JUDSON: That is incorrect. This was the -- I'll just tell you a little bit -- maybe this is good context for the 1,800 chemicals.

The first 300 were all the pesticide actives which we tested because we had good in vivo data for. And again, this whole effort was not done to do EDSP. We're doing broad screening. All the pesticide actives, and in the next thousand were -- as many more chemicals as we could find that had in vivo data of any kind, as well as some pharmaceuticals as good references.

The second 800 were almost all chemicals taken from the EDSP list, the 10,000 list. We tried to get down as deep as that. But there was no selection for things that we thought might be ER or AR active.

DR. JAMES MCMANAMAN: Thank you.

DR. SCOTT BELCHER: Could you comment about the sensitivity of your confirmation assay? Your R1881 in the high dose is probably about 50-fold higher than --

DR. RICHARD JUDSON: Something like that.

DR. SCOTT BELCHER: There is some concern that this confirmation assay is going to be so insensitive that you're going to be pushing it with your confirmation flags that you have.

DR. RICHARD JUDSON: Right, and truth in lending here or something. The high dose was actually an accident and so, it was then pointed out to us that, oh, you shouldn't have tested so high. So we went back and tested it low, but realized, okay, this was actually kind of a cool -- we could do a cool experiment, given those two concentrations. The high

dose is not the ideal high dose concentration to use if we were planning this a priori.

DR. SCOTT BELCHER: Just to follow up, I would agree that it's not a good experiment or a good approach to be using and that decrease in sensitivity may be problematic.

DR. RICHARD JUDSON: Right, but having recognized this was a good thing to do if you do it right, all of the new antagonist assays being run at NCGC are being run with two concentrations -- not necessarily quite as insensitive -- the high is not quite as insensitive as we've run here.

DR. SCOTT BELCHER: We still have the problem that only the really strong antagonist would have the three points, I think in your scale, and then you would be losing sensitivity in that.

DR. RICHARD JUDSON: No, so you'll just have two points. You have the high dose and the low dose. AC50 is for antagonism. It's the ones in the middle that are -- actually it's the ones that are really potent and in the middle. The only ones you lose are the ones that are really weak to begin with.

DR. KRISTI PULLEN FEDINICK: I had a question also about the chemical space. You're only

looking at chemicals that are soluble in DMSO, is that right?

DR. RICHARD JUDSON: Yes, that is a limitation. They can't be volatile, yes.

DR. KRISTI PULLEN FEDINICK: So, then chemicals, say, that would fall under the Safe Drinking Water Act that might be soluble in only water, how do you think that this model would work for those types of chemicals if you're only testing DMSO soluble ones? Do you think it's applicable to chemicals that would fall outside of the limitations of ToxCast?

DR. RICHARD JUDSON: That's a broad question. A lot of these chemicals, a lot of them are water soluble. They have to be DMSO soluble, but that doesn't exclude being water soluble.

Chemicals which are not DMSO soluble, we have talked for years about having a bunch of these assays run in water. Some of them we've at least done a little bit of testing, but it just hasn't been a high enough priority to do. In principle, there is no reason you can't do that.

DR. KRISTI PULLEN FEDINICK: Can I just follow up? But the current assay does allow for water

solubility, alcohol solubility, and DMSO solubility? The current EDSP Tier 1 test?

DR. RICHARD JUDSON: Yes, right.

DR. KRISTI PULLEN FEDINICK: You could do all the solubility? So ToxCast is more limited than the current assay?

DR. RICHARD JUDSON: Current implementation assay, yes.

DR. KRISTI PULLEN FEDINICK: Great, thank you.

DR. THOMAS ZOELLER: To follow up on Dr. Ehrich's point, the way you do antagonist assay is have a single dose of an agonist and multiple doses of test chemical?

DR. RICHARD JUDSON: Correct.

DR. THOMAS ZOELLER: If that test chemical has an allosteric mechanism, if it's a non-competitive, let's say, inhibitor, you could predict what that would look like. It wouldn't look like a traditional kind of competitive inhibitor. It might be useful to look at those data sets that are kind of weird and ambiguous; to flip that around and do a single concentration of test chemical and

multiple concentrations of agonist. It changes the affinity basically.

DR. RICHARD JUDSON: Right, and I don't know whether that's part of the mandate of the current SAP and this meeting. I would certainly be interested in hearing some chemicals that we would expect to behave like that, yes.

DR. GRANT WELLER: I just had a question on your model fitting procedure. It appears that you're fitting different -- you're estimating receptor values at different concentrations of a given chemical independently. As just a non-biology expert, I'm wondering is that -- what's the reason for doing that as opposed to doing some kind of maybe smoothing over different concentrations. Can you comment on that?

DR. RICHARD JUDSON: A technical detail. Every one of these different assays was run at different concentrations. Some did an eight-point concentration response, some did a twelve-point, some did a six-point, and they didn't all line up. I actually had to do a smoothing before that. I took, for each assay, the AC50s, the hill slope, and the top. Some people at least understand those.

You have the actual assay data which is somewhat noisy, and you have this smooth curve going through that. Then what we did was we interpolated, we chose a set of 12 or 14 concentrations and we interpolated to get a smooth version of the curve that goes for each of those concentrations.

The real answer to your question, this was a simple way to do it. Do it one concentration at a time, these interpolated concentrations, and then knit things back together again. I'm sure there would be some more sophisticated way that sort of does it all at once, but -- you know.

DR. REBECCA CLEWELL: My question -- and you may be getting to it, so I apologize if I'm rushing you. What I was wondering is, after you've done this confidence scoring, how is that applied to your decision-making context in terms of if this is or is not androgen active?

DR. RICHARD JUDSON: Yes, why don't you hold that, and there is at least a chance I'll answer that.

DR. REBECCA CLEWELL: Okay.

DR. JAMES MCMANAMAN: I'm just looking over the next set of slides and I'm wondering -- we're

scheduled for a break in about a half an hour. It looks like that there is a break and then there is a very short presentation by Dr. Bever, and then we break for lunch. I'm wondering, does anyone have a strong feeling one way or another, if we should break now and then go for the next set? Does that make sense?

We'll have a 15-minute break now and then we'll continue on. It seemed like a good place to do this.

[BREAK]

DR. JAMES MCMANAMAN: Welcome back. Before we move on, the statisticians and the mathematicians in the group have some clarification questions about the models. It seems an appropriate time to do that, if that's okay. Go ahead.

DR. VERONICA BERROCAL: Thank you. I guess we -- at least I felt that it was not maybe an important question to ask before, but now I really would like to understand what are the specifics of this model. I am confused about slide 48.

DR. RICHARD JUDSON: Let's see if I can figure out how to get to slide --

DR. VERONICA BERROCAL: It's just on the next slide.

DR. RICHARD JUDSON: The next slide, okay, yes.

DR. VERONICA BERROCAL: I'm just confused about these curves that I -- these two panels. So, the panel on the left it says, assay data. I counted eight curves?

DR. RICHARD JUDSON: I wish I had my reading glasses on.

DR. VERONICA BERROCAL: I guess I'm just confused about -- so, do each of these curves correspond to a different node?

DR. RICHARD JUDSON: No.

DR. VERONICA BERROCAL: Yes, I guess I'm just confused about this.

DR. RICHARD JUDSON: So, each curve corresponds to a different assay.

DR. VERONICA BERROCAL: But then there are 11 assays and I only see --

DR. RICHARD JUDSON: Right, so they're not all on. Notice that the -- we test up to 100 micromolar, and so as we're starting to shift -- maybe

some of them are not active, but let's go back to this slide.

For an antagonist, we should light up, one, two, three, four, five, six, seven. So a true antagonist should light up the purple ones here, only seven. A true agonist should light up, one, two, three, four, five, six, seven, eight, nine. So that's clear?

DR. VERONICA BERROCAL: Right. And so the N receptor will be two. In the previous slide that you have here, X equal to N receptor and receptor is two.

DR. RICHARD JUDSON: Sorry?

DR. VERONICA BERROCAL: So, in that equation where it says, X equal to sum from one to N receptor.

DR. RICHARD JUDSON: Right.

DR. VERONICA BERROCAL: The N receptor would be just two. You have R1 and R2.

DR. RICHARD JUDSON: No, so the N receptor is --

DR. VERONICA BERROCAL: Either five or nine.

DR. RICHARD JUDSON: It's actually -- so, there is one, two, three, four, five, six, seven, and there is an eight, a nine, a ten, and then there are a bunch of others. So, the number of these pseudo-receptors or nodes is -- there is one for each group of assays, and then there is one for each assay individually.

DR. VERONICA BERROCAL: I see. And then I guess the other question -- again back on slide 48 -- is once these individual assay curves are generated, how do you get to the two curves that you see on the right panel?

DR. RICHARD JUDSON: Right. Again, this is the assay data. And recall, these are these interpolated, smoothed curves, okay? What the model goes through and it's testing -- sorry, to go back here, there are 15 or 16 of these R's. We're actually developing a -- the computer is guessing the value of each concentration, the value for those, call it 16, of those R's. Then it finally, at the end, says, most of those get driven down to zero.

What we show here are actually all 16 of those curves are there, but most of them are just sitting at zero. The only two that have any weight

are the two that are shown here. Again, if we didn't use the penalty term, you would see more of those kind of hovering close to, but not equal, to zero. So, that penalty term drives most of those zeros down to zero.

DR. VERONICA BERROCAL: Thank you.

DR. MICHAEL PENNELL: Maybe I'm getting this a little slower. Could you go back to the previous slide and kind of just walk through some of the systems of equations, like A1, A2. Just help me understand exactly what the system is that you're solving.

DR. RICHARD JUDSON: Do you all have your hard copies of the slides?

DR. MICHAEL PENNELL: Yes.

DR. RICHARD JUDSON: Pull up the hard copy of the slide that has all of those assays listed. Pull up slide 41. I won't go back to that. If you care, look at that.

The first column is the number. A1 is the first assay, A2 is the second assay, and so on. We have 11 As, and we have -- again, I don't remember -- I think it's 16 Rs. Computationally -- and this is using a constrained gradient optimization method, so

it's R, it's optim -- X, you know, if you want to know real details. So, you give it this set of equations, and then the F matrix is that connectivity matrix.

The penalty function, it's simply, you hand it -- the true -- you're doing this all one concentration each time. The true As for those 11 As, then it uses its current value of the 16 Rs. It does this matrix multiplied to get the predicted As. And then given the Rs, it does this penalty -- calculates the penalty term from the R vector. You may be trying to make more out of this than there is.

DR. MICHAEL PENNELL: I think I'm still missing this connectivity matrix.

DR. RICHARD JUDSON: Look at the picture. So, the connectivity matrix is if there is an arrow -- one of these black arrows -- between a node and an assay, then F is one. If there isn't a direct line, F is zero.

For instance, between -- this is a good example. Between this node and this assay, there is no connection. The F would be zero between those two. Whereas between this node and these assays, F is one.

DR. MICHAEL PENNELL: I think why I'm confused is -- do the Ns here correspond to what

you're calling Rs in the equation? I mean, I'm a little confused with the notation, I guess.

DR. RICHARD JUDSON: Yes. I'm sorry. Again, this is something that -- it's this committee's fault, right? The last time when we did the estrogen receptor, I called all of these pseudo-receptors, and so they were all Rs. You guys complained about it, and so they became nodes. The problem is, some of the terminology didn't go from R to N. N and R are the same thing. Sorry.

DR. IOANNIS ANDROULAKIS: Just one quick, maybe two technical questions. I'm trying to understand, so basically what you're doing is you're solving this problem for different doses and the hope is that as you increase the dose you should start seeing more of these R stars lighting up? Because then you go down the pathway?

DR. RICHARD JUDSON: Right.

DR. IOANNIS ANDROULAKIS: Then at the same time, you're solving a very highly non-convex problem. And as you say, you're sort of penalizing, so there is no particular reason why certain nodes should be consistently on or off.

Out of curiosity, how do you maintain this sort of continuity? As you move along the doses, how do you make sure that the way these Rs light up is because of the dose sort of pushing the pathway down, as opposed to hitting different minima?

DR. RICHARD JUDSON: There are actually two answers to that. First, if we go back to this pathway, there is no biological expectation that low doses hit this and then as you get high doses, you hit this. All of these should turn on at exactly the same concentration. The fact that these assays don't turn on at the same time, that's just an assay technology issue. Some assays are just more sensitive than others. There is no correlation between when this turns on and going down the pathway.

But then you say, why on earth is this red curve smooth? It turns out, the model is -- essentially we're modeling in a very simple way the real biology and the real biology is smooth. I was pleasantly surprised to see that those curves turned out to be smooth. The model would have failed if that hadn't been true. All we're doing is really essentially doing a fancy averaging of these smooth curves, and so the result has to be smooth.

DR. JAMES MCMANAMAN: All right. That was Dr. Androulakis' question and Dr. Judson answering all of those questions.

I think if we're ready to go, I think we should move onto the next -- Dr. Pennell?

DR. MICHAEL PENNELL: I have one more quick question. You didn't give the expression for the z-score, but what I'm seeing in the white paper kind of confuses me. It's just a positive/negative issue.

If it's something that's not -- if the AC50 is significantly both cytotoxic range, you should get a highly negative z-score, right?

DR. RICHARD JUDSON: Correct.

DR. MICHAEL PENNELL: If you're taking the log AC50 minus the log AC50 for cytotoxicity, right?

DR. RICHARD JUDSON: Right. If you look here, these assays have negative z-scores. Z-scores can be positive or negative. All we care about is, are you really positive, are you zero or worse?

That obviously didn't answer the question.

DR. MICHAEL PENNELL: Hold on. I don't think I announced my name. This is Michael Pennell from Ohio State again.

If you only care about the high z-score -- so, the way the methodology is written, is you're flagging ones that have a z-score of greater than three. And you're saying that that has a cytotoxicity -- excuse me, has an activity which is below the expected range for cytotoxicity, right?

DR. RICHARD JUDSON: Correct.

DR. MICHAEL PENNELL: It's just a computation issue, right? If you're taking the log AC50 for the chemical, minus the log AC50 for cytotoxicity, you should get a negative value, not a positive value?

DR. RICHARD JUDSON: Okay, you're right.

DR. MICHAEL PENNELL: It's like that in your publication too and it really confused me.

DR. RICHARD JUDSON: There should be a negative -- there is a missing negative sign, sorry.

DR. MICHAEL PENNELL: I just wanted to clarify.

DR. JAMES MCMANAMAN: If there aren't any other questions, I think we're ready to move on.

DR. RICHARD JUDSON: I hope everybody has seen heat maps before. Here we have -- these are actually the nodes. They're not pseudo-receptors, they're nodes. There is -- one, two, three, four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fourteen, fifteen, sixteen -- there's seventeen. I guessed wrong.

So, there is the agonist, which is R1, the antagonist, which is R2. R3 through R7, which are these nodes that have to do with specific groups of assays, and then there is a pseudo-receptor or node for each of the individual assays. Chemicals are on this axis, and we're only showing the 763 chemicals that are active in at least one assay. So, there are another 1,100 which are just blank.

The color dark red has a high AUC, a light red has a low AUC. This is the band -- these are the agonists. There aren't a lot of them. There are some that are really, really potent down here. There are a few up here that have some activity. It could be an agonist or it could be an antagonist. And again, remember whichever one has the highest AUC

wins. Those are the agonists. Then the antagonists are mostly this band. There are a lot of antagonists, a lot of which are not true. That's why we are going to use the confidence score to try and filter those.

Then if we look at a given set of chemicals to think about assay interference, so this is a set of chemicals which is only active in assay A11, so that transactivation assay. The AUC all goes to that assay and so we can say, no, because they're active in that assay does not mean they're androgenic, that's just false activity.

Then there are these chemicals where they are active -- they have some potency in both, and again you can't really see it very well, but this particular chemical could be an A11, or it could be an antagonist. So, it's obviously active in some other assays, but if you actually look at it, it's darker red in A11 than in antagonist. That's a chemical which the model has said, you know what, we think it's an assay interference and it's not a true antagonist. It would move over into that column.

You've seen some of these curves before, but this is an example of a true agonist. We haven't seen any agonists before. Testosterone

propionate, the assays are quite potent, but notice again, there is a lot of variability in the potency of the assays, but all the assays are turning on well below where cytotoxicity happens. You can just barely see the agonist curve, the agonist consensus curve, this blue thing up here.

Then this is a true antagonist. Again, there is variability in the assay potency and then you get this red curve for the antagonist. Then there is this alternative, some other mode, which has some probability. But it goes up and it comes back down again.

Here is an example of a narrow assay interference. This is PFOS and it's actually lighting up two of the three -- these radioligand binding assays. But it doesn't light up anything. It's very clear that there is some -- somehow that -- I can make guesses at what PFOS is doing to those cell-free assays. It's interfering somehow with that. But it only interferes with those assays, because PFOS is able to get right to the receptor, whereas maybe it can't even get into the cells.

This is this confidence scoring that can, again, go from negative one to six. This is

looking at a correlation between the confidence score on the X-axis, and the antagonist AUC on the Y-axis. And so you see things that have high AUC also tend to have high confidence scores. But what we really are using the confidence score is to deal with these issues, things which have really high AUCs, but really have low confidence scores. These are mainly the ones that are going to end up getting filtered out.

Apologies. I'm shifting gears here. This is a new concept, but something that was asked for last time around to quantify uncertainty. We don't really use this anyplace here, but just to point out that we've actually done this and are thinking about how we would incorporate this.

The basic approach is, with the model I told you before, for each chemical we have 11 assays and each of those assays has six or twelve concentration points at each concentration. We have two or three replicates. We fit the curves and then we build the model from that.

Here what we do is we say, okay, there's really -- if you only had five of the six concentrations, you only had that data and you fit the curve, you would get a different AC50. You might even

decide it was not a hit. You're kind of borderline. Maybe it goes from a hit to a non-hit.

What we did here, it's kind of the simplest -- what I think is the simplest way to do uncertainty quantification -- is for every curve, for every chemical, we do a bootstrap. It means that we draw 80 percent of the data for the chemical, and we refit the curve. And we do that a thousand times. Now for every chemical, every assay, we have a thousand different values of the AC50 and is it a hit or not a hit.

Then for those thousand replicates, we go rerun the whole model. And so now, the AUC changes as you change the underlying data. We actually have published -- we have a publication under review at a journal on the ER model. This will get published at some point. This is just an illustration of what you get out of this. This is the AUC, the order of the initial model. And it's whatever you actually have, the black dots with the antagonist and orange dots for the agonist. You will have both a black dot and an orange dot for each chemical. We have many more antagonists -- all the antagonists, these black ones. And for the most part, the agonists are down here

below. We can see the confidence intervals around the point estimate.

One comment we got the last time is, I don't like that 0.1; be more precise about the 0.1 cutoff. What this shows is it's somewhere around 0.1, because the whole uncertainty of the model can be relatively large. Another way you could add confidence to your score is you say, okay, is the integral over the confidence interval, is it mostly above 0.1 or mostly below 0.1? So you could eliminate some of these kind of borderline ones, because it only peaks above 0.1 for most of your thousand replicates. It's a technique we've developed and I can talk about more at another venue.

Another comment that came up from the previous committee, and two public commenters are very concerned about this AUC versus potency. So, this was not in the white paper, but it was important enough to point this out. People say, you give me an AUC, but I can't use that in a quantitative risk assessment because that's not potency. But our response to that is the AUC and the confidence scores are being used in the prioritization.

It's really saying, do we believe this chemical is active against the androgen receptor or not? Once you decide, yes it is, then you just go back to the raw data and you can figure out what the potency is for your quantitative risk assessment. That's really the main approach to that. All of that data is publicly available. The potency values for all of the assays are in the supplemental data.

But the public comments had shown, oh, there is no correlation, but I had to go back and redo the arithmetic and it turns out -- this has to be true -- that there is a strong linear relationship between the AUC and the average AC50 for the assays. By construction that has to be true, because the AC50 is really kind of naturally on a log scale. So, you have to compare the log of the AC50s with the AUC.

The antagonist, you still see this linear relationship. There is this group of chemicals down here which have -- and notice the coloring is really confident ones, with high confidence score, are red; this intermediate, kind of iffy, is blue; and then the really low confidence, you don't believe, are white. Then interference ones are chemicals that have an AUC for antagonism, but some other node has a

higher AUC. So we wouldn't call that an antagonist at all. Most of these that are off diagonal are either some other node wins, or they're really low confidence score.

If you zoom in on this area close to the AUC of 0.1 -- I haven't quantified this, but the number of reds, highly confident ones, is in the small minority. There are lots of blues, whites, and blacks there. This really addresses the issue of AUC and potency really are correlated. But if you want potency, you have to go back to the raw data.

Now we're actually getting to the meat of the issue and addressing the charge question, is this approach good enough to actually be used as an alternative in Tier 1.

What we're looking at here is -- let's focus on the agonist for the moment. This is the AUC. These are all of the reference chemicals from -- it seems like hours ago, we talked about with their potency. A circle is green if the reference set said it's positive, and it's red if the reference set said it's negative. I specifically called out this ambiguous region.

We see most of the greens -- in fact, all of the greens are well above 0.1, so there are no false negatives. There are a couple of ones the literature said are negative -- one that's relatively high potency on the AUC scale, and one that's in this ambiguous region. On the antagonist, almost all of the literature positives are positive, well above 0.1. There are a couple in the ambiguous region. Then there is one in the -- that's called negative. It's important to dig in and say, why are we missing the things we're missing? I'll go into that in a bit.

Looking at the statistics, predictivity statistics, and focusing on the balanced accuracy, if we put the ambiguous chemicals into the positive class -- sort of err on the side of caution -- then we get the balanced accuracy of 0.95/0.97. Actually, very high and probably higher than most of the -- at least as high as typical assays that we validate.

If you put the ambiguous chemicals into the negative class, still it's -- 93 percent is the lowest balanced accuracy. So the statistical support for this model against these well-studied reference chemicals is quite high.

Then we did a comparison -- not necessarily -- I don't remember what the charge question says. These things look like -- we do a good job of predicting true and positive androgens. But what we really want to know is how it compares with the AR binding assay, the Tier 1 binding assay. There were 101 chemicals that had been run in this one assay, not a bunch of different assays, during the assay validation. Fifty-five were run actually -- they were compiled by ICCVAM -- they were run in multiple labs. There is good evidence for the activity of these -- I'll say that and then I'll caveat that later on -- for these 55 chemicals. These are kind of a reference set for the AR binding.

During the List 1, there was an initial set of chemicals that this group, the EDSP, had people run. Those are the List 1 chemicals. There were 47 chemicals actually run in this assay by commercial labs. But they were only run once. One lab took this chemical and ran it once.

This is the comparison there. Let's look at the ICCVAM first. There were 24 actives that ICCVAM had in their list. Twenty-two of those we called active. A binding assay will be active whether

you're an agonist or an antagonist, so we can't distinguish there. Twenty-two out of the twenty-four are positive, and one was inconclusive and one negative. I'll go in and drill down to what is going on with those two.

The ICCVAM inactive, there were 31 -- we only called 19 of those negative, and so there were a bunch that were called positive. That's an important set to understand. There appears to be lots of false positives, but I'm going to argue that's not really true.

Then the EPA List 1 chemicals, there were nine actives. We called almost all of those inactive. We missed all of those, so that's worrisome. I'm not going to worry about the inconclusives. Then there were 31 inactives on the List 1. We got 24 of those and then -- my arithmetic doesn't quite work here. Anyway, there were four that we missed. I think that's four and this should be whatever the balance is, yes. Twenty-seven, right. I believe that's true.

So let's look at those discrepancies. These are the ICCVAM actives and the ones that we called inactive. There were these two: there is

atrazine and methoxychlor. And if you go back and look at the ICCVAM activity, they were both active at very high concentrations. Atrazine is at 53 micromolar and methoxychlor at 185. We only test up to 100 and chemicals that are even close to 100 micromolar, it's kind of iffy whether we get that. So these are just outside of the boundary for us to be able to really detect that with our standard 100 micromolar protocol.

The ICCVAM inactive, that we called active -- there were a bunch of them. There were seven antagonists, two agonists, and then a couple of inconclusives. And notice, I've just underlined a bunch of them. Nine out of those twelve were actually estrogenic. There is a good -- it makes sense that chemicals that are estrogens will at least be weakly active in the androgen receptors. There is this cross-talk which is not surprising.

A little bird told me to go actually read this ICCVAM document where they list those chemicals, and it turned out a bunch of those estrogens were never tested. The expert panel said, oh, they're estrogen so they can't be androgen, so you can just use those as your negative reference

chemicals. A bunch of these reference negatives in the ICCVAM panel were never actually tested for their negativity. So I don't know what the answer to that is.

Discrepancy with the EDSP List 1. The List 1 active and we called inactive, there were a bunch of those. Almost all of those were active at very high concentrations. The Tier 1 AR binding assay protocol says go up to one millimolar, so 10 times higher than we do. Almost all of those were outside our testing range, with the exception of phosmet.

Then the inactive List 1 where we call active, we can have -- so, it was actually six -- it should have been another number on the other slide, six. Most of these have low confidence scores. Most of these are antagonist, but confidence score of one. We would say, you know what, those are not really positive anyway. We would filter those out. Most of our false positives we would throw away.

To summarize all of that, so the multi-lab chemicals that have been tested a lot; We got 22 out of the 24 actives, and the majority of these that we called active that they called inactive, were these estrogens where they had never actually done any

testing to confirm they were inactive. And in the single lab results for List 1, most of the actives we missed were -- all but one of them were very high concentration, so we couldn't detect those. I don't know where the 1,711 came from. Nonetheless, most of the false positives have this very low confidence score.

So, that's the results. I just have a little bit of discussion. Do people want to ask questions here before I go on?

DR. JAMES MCMANAMAN: Dr. Zoeller I think does.

DR. RICHARD JUDSON: Sure.

DR. THOMAS ZOELLER: In kind of thinking about this issue of false negative and false positive, you have many assays to determine whether a positive is a true or a false positive. But when it comes to a false negative, it's really that you've got a reference chemical that was negative in ICCVAM or some other kind of setting, or that was positive that shows up negative in your assay. So then it's a false negative.

But when you go to apply this to chemicals that haven't been tested before, there is no

way -- it seems to me -- to discriminate between false negative and true negative. When I think about confidence, it's almost one tail.

DR. RICHARD JUDSON: You're addressing the issue of false negatives. We run all of our 11 assays and we conclude through the model that either we don't see any activity in any assay, or we conclude whatever activity we see is some interference mode. That's really -- and how do we know that those are really negative? We don't. That's the basic answer.

But it finally goes back to, what are we doing this for? Our basic approach is we're trying to prioritize. So we have plenty of chemicals for somebody to look at that are clearly positive. A lot more than anyone has ever sort of done any deep dive on. And those that are in the middle -- and we have even more in the middle. Then we have these things that are really negative. So that's one answer.

The other answer is, okay, with the assay we're trying to yield an alternative for, they have one assay. If it's negative in that one assay, no one is ever going to go look at it again. At least we've gotten 11 up to bat or something like --

whatever your favorite metaphor is -- rather than just one.

DR. KRISTI PULLEN FEDINICK: Can you go above 100 micromolar for testing in ToxCast or is that a technical limitation?

DR. RICHARD JUDSON: It's a practical limitation. Again, we're ultimately -- between us and NCGC, we've tested about 9,000 unique substances. To be perfect, what you would do is, every single chemical, you would figure out what the limit of solubility is in your favorite media, and then you would test up to the limit of solubility. But that's really, really hard.

From a practical standpoint we said, you would have to eat a lot of stuff to get to 100 micromolar. We sort of -- one can argue with this, but the kind of accidental, non-intentional exposures that we're worrying about -- we worry about for the most part, it's hard to imagine how you get above 100 micromolar. So that's sort of a good upper limit. And then we actually say, okay, is it soluble at 100 micromolar, will it solubilize everything? And if it's not, we actually go down by a factor of two or

ten or something like that and try again. It really is practical.

DR. KRISTI PULLEN FEDINICK: But for say, worker exposure, where it's not just consuming, it's actually being in the presence of that chemical at high concentrations?

DR. RICHARD JUDSON: Correct.

DR. KRISTI PULLEN FEDINICK: For the chemicals that you missed, in particular. Is it possible for that 100 micromolar limit to be expanded? Is there a cost limitation to that, or why -- could you in theory -- not in theory, but could you in practice, go above that 100 micromolar per chemical set?

DR. RICHARD JUDSON: We could go up to the limit of solubility, but you can't do that for thousands of chemicals at a reasonable cost, no.

DR. REBECCA CLEWELL: Can we go back to where you showed the confidence scoring on the plot? The sort of box and whisker?

DR. RICHARD JUDSON: Yes.

DR. REBECCA CLEWELL: No, the big boxes. The box plot, yes.

DR. RICHARD JUDSON: This one?

DR. REBECCA CLEWELL: Way back, yes.

DR. RICHARD JUDSON: Right.

DR. REBECCA CLEWELL: I was wondering if you could maybe explain or provide an example of a situation where you would get -- and clearly it happens a lot -- but where you have the sort of red box outline here. Where you have a very high AUC, but a very low confidence score.

I'm trying to put this model together in my head, and in order to do the kind of combined AUC score, I'm assuming that you normalized everything so that they have a similar activity. Because these different assays can have very different sorts of dynamic ranges.

DR. RICHARD JUDSON: Right, yes.

DR. REBECCA CLEWELL: Then so, I'm trying to picture a situation -- because the only thing in my mind that would make that work, to have a very high AUC and a very low confidence score, would be having a very high activity in one single assay.

DR. RICHARD JUDSON: No, actually it's not that way. Think back to this burst idea. If you have a chemical that causes three or four or five of the assays to light up purely due to cytotoxicity, or

cell stress, whatever that is. So you have all the assays on, and so it's going to finally say -- and you have assays for multiple technologies that are on. It's going to give you a high AUC.

But then the confidence score will say, you know what, you don't have the shift in the right direction, so you lose points there. Everything is in the cytotoxicity region. You lose points there. That is how you would get that situation.

Most of these -- I would have to look at individually, but my guess is most of these are lighting up at multiple assays, multiple technologies, in the cytotoxicity region.

DR. REBECCA CLEWELL: Do you think that the major driver then is probably the z-score for the confidence interval, generally?

DR. RICHARD JUDSON: Yes. Well, you get a lot of -- remember -- I won't go back to that chart way back, but you get a lot of points for getting the shift in the right direction. You already lose three points if the shift is not the right – or you could lose four points. You could actually get a negative value. And then if you have the -- you're all in the z-score, you lose points too.

DR. REBECCA CLEWELL: I have one more question. Is that all right?

DR. RICHARD JUDSON: Yes.

DR. REBECCA CLEWELL: Then when you go forward and you start doing a kind of positive -- true positive and false positive, true negative and false negative -- there was sort of a lot of information there, so it can be kind of hard to put it all together in my mind.

What I'm wondering is, for these ones where there is a clear kind of reason for -- first of all, when you did the false positive and false negative and you did the hit calls, does this include your confidence interval consideration?

DR. RICHARD JUDSON: It does not.

DR. REBECCA CLEWELL: It was just the AUC score?

DR. RICHARD JUDSON: This is just AUC. And then all of those next slides that have all of those words on them, words and numbers and complicated -- those are where we try to explain, why do we have a false positive? Is there a good explanation that has to do with the confidence interval?

One thing we haven't done here is to say, I'm the expert, so if you have an AUC of 0.1 and the confidence interval is two or less, I'm going to call it a negative.

I think that is something that -- it has to be a wider discussion. Certainly within -- the program office would have to come up with the final decision of how to do that with input from you folks, potentially.

DR. REBECCA CLEWELL: It's hard to see doing much better than, like, 90 plus percent balanced accuracy. What would be interesting and, I think, informative, is if we could compare what happens if we use different confidence score cutoffs, and how that changes your predictions here.

DR. RICHARD JUDSON: Yes.

DR. REBECCA CLEWELL: Also, if they were never tested and they're not actually shown to be androgen negative, then I don't know, maybe they shouldn't be used as a control.

DR. RICHARD JUDSON: I got these slides -- so, Dr. Kleinstreuer actually wrote the -- she was the primary author of the AR paper. I certainly helped. And she put these slides together. Only at

the last minute as we were getting ready she said, oh, you should go look at that. Probably if we were to publish this, we would go back and scrub all of those that actually are not supported by data.

DR. KRISTI PULLEN FEDINICK: I had another question. We couldn't necessarily apply the accuracy sensitivity or specificity to chemicals that aren't DMSO soluble? Is that right? So, we couldn't know if this is applicable across all 10,000 chemicals for EDSP? We would only know if they were applicable over the 1,800 or so chemicals within ToxCast?

DR. RICHARD JUDSON: Well, there are two questions there. One is, we have data for 1,800 chemicals here. We actually have more data back at home, but for this -- we can't make any statement about a new chemical without testing it. Dr. Thomas reminded me that he told me before -- we actually are testing -- we have a 96 or 384 chemicals in water that we're actually testing right now in an alternative assay. We actually are moving in that direction.

DR. KRISTI PULLEN FEDINICK: But just to clarify. So, we wouldn't know whether or not these tests would be accurate for chemicals in water or not? We couldn't apply this -- we couldn't look at the

balanced accuracy of 95 percent and say that would also then apply for chemicals that are water soluble?

DR. RICHARD JUDSON: I guess I wouldn't know any scientific reason why -- ultimately, the well that you're testing in is watery, right? So, the only reason you -- you just need to get stuff out of powder into solution with the DMSO before you put it back into water. I would think there is no a priori reason why it won't work perfectly well for chemicals soluble in water.

DR. KRISTI PULLEN FEDINICK: Or volatility, not just water.

DR. RICHARD JUDSON: Yes, volatility is a whole different issue, right.

DR. KRISTI PULLEN FEDINICK: So, we don't know that those -- thank you.

DR. EDWARD PERKINS: Back on slide 62, uncertainty quantifications. Using bootstrap replicates for the concentration response curves, you were looking at distribution of fit parameters and model selections. Did you find any influence of model selections for curves and so forth on that? On your AUCs or how that might affect -- or am I confusing things?

DR. RICHARD JUDSON: A gory detail that Dr. Perkins probably knows about. For every chemical assay, we fit three curves. We fit a flat curve. Does it look flat? Does it look like a hill curve? Or does it look like a gain/loss, up and down again?

As you subtract points away, you may go from saying it was initially a hill, but it either goes flat or it goes to a gain/loss. In this modeling approach, we actually let it go to whatever it was going to go to. Specifically, I'd have to dig deep to answer your specific question.

DR. EDWARD PERKINS: Yes, because in the pipeline you're scoring -- only the hill or the gain/loss are scored as active, right?

DR. RICHARD JUDSON: Correct, right.

DR. EDWARD PERKINS: So, the idea is that those generally fit most all curves that you're seeing in all the assays?

DR. RICHARD JUDSON: People have asked, well, why didn't you fit exponentials and other kind of curves, and -- just because we don't. We can have a long discussion about that, if that's what you're asking.

DR. EDWARD PERKINS: Thanks.

DR. J. DAVID FURLOW: Just one quick question. I'm also thinking of the large number of antagonists that are showing up in a lot of these assays and how to filter them and assign confidence scores to them, right?

So looking at seeing cycloheximide show up, right? It's almost a test case that in that case, that should score low because it should go the wrong direction in an agonist assay. Isn't that right? So if you have some activity --

DR. RICHARD JUDSON: It shouldn't go in the right direction.

DR. J. DAVID FURLOW: If you're testing it without any androgen alone, right -- or around, it should go down?

DR. RICHARD JUDSON: Right.

DR. J. DAVID FURLOW: Because there is some activity, you're going to inhibit the reporter, right?

DR. RICHARD JUDSON: Right.

DR. J. DAVID FURLOW: So, if those are parallel, right? So when you add androgen and you don't have androgen, they both go down, that should decrease its confidence. Is that captured in this?

DR. RICHARD JUDSON: Yes. If I looked at the confidence score for cycloheximide -- was it on one of the slides? It should have a low confidence score.

DR. J. DAVID FURLOW: Right.

DR. RICHARD JUDSON: Because all the activity -- cycloheximide is so cytotoxic, almost all the activity should be in the cytotoxicity region too.

DR. J. DAVID FURLOW: It should be. In some of our assays we've seen cycloheximide going down well before there is cytotoxicity, for reasons I don't understand. It depends on the nature of the reporter. But it was in the Class B, right, that it was air pathway active, but ICCVAM inactive.

DR. RICHARD JUDSON: Again, we've studied the cycloheximide. We've had big arguments about that and it's complicated.

DR. J. DAVID FURLOW: I mean, it should show up in all of the assays across B53. Whatever your tox 21 readouts are, right? Cycloheximide should be showing up in a lot of those.

DR. RICHARD JUDSON: It does.

DR. J. DAVID FURLOW: It may be well before -- for whatever reason, before there is toxicity.

DR. RICHARD JUDSON: Right.

DR. KRISTI PULLEN FEDINICK: Just another quick follow-up question. For chemicals in that confidence score equal to one, have you gone back and looked at whether or not those would be potentially false negatives? Getting at Professor Zoeller's comment earlier. Just having looked through a little bit -- in preparation for this meeting, there is one chemical, Mancozeb for example, that would have a confidence score of one and would have -- it was an antagonist score of 0.0517. But it was a chemical that you had flagged in 2010 as being active across a number of assays in the initial ToxCast, just kind of released to the public.

This is one that would potentially have a DMSO solubility issue. It could also have non-classical binding. It's a known antagonist, but it's showing up with a confidence score equal to one. Is there a way to really know whether or not that one cutoff -- or understanding what's in that one pool, so

to speak, so that we can understand what's really happening there?

DR. RICHARD JUDSON: Mancozeb, which is a zinc -- zinc manganese. It's a real nasty thing which is very cytotoxic. It turns on all sorts of assays, and so my conclusion is, it's just causing the cells to be really unhappy and all sorts of stuff is turning on.

So I would say it has a low confidence score because all of the activity is in -- it doesn't push it in the right direction and the activity is mostly in the cytotoxicity region.

Now, you made a comment that is a known antagonist. I would be interested in seeing the literature that proves that.

DR. KRISTI PULLEN FEDINICK: In CHO-K1 cells, for example, there is a clear antagonist binding that if you're looking at certain concentrations -- not even certain concentrations, but there are curves that will demonstrate absolutely that it binds to the AR receptor. But we can talk. I'm happy to share references too.

DR. JAMES MCMANAMAN: I think we should move on.

DR. RICHARD JUDSON: I'm almost done. Discussion, just summarizing. I'll just quickly summarize what we've talked about. We ran these 11 assays in 1,855 chemicals to sort out whether antagonists are true or false. We have this AUC value that we've combined with these confidence scores. Just a summary, so 1,100 chemicals were inactive in all the assays. We could argue about whether we have 11 false negatives for all of those chemicals, but I would argue probably not. Those are probably really not androgenic.

Five hundred sixty-two were active in at least one assay, but were not classified as either agonist or antagonist. We have 33 agonists and a lot of -- still almost 200 antagonists, but 140 of those have a confidence score of three or higher. My rough rule of thumb is it's got to be three before you start paying attention to it.

People have looked at this and said that's way too many antagonists. But if you look at the chemical structures for those 146, most of them fall into a relatively small number of chemical classes. They're not random chemicals.

We have a bunch of steroids, a bunch of bisphenols, relatives of Bisphenol A, a lot of chlorophenols -- they look like, for those of you who know Bisphenol A. It's two benzene rings with a methyl group in the middle and you have a hydroxyl group, so those are the bisphenols. If you replace the hydroxyl group with chlorines, that's another common class that -- it makes sense chemically to me that those would be binders. And then Bianiline, so you actually have -- instead of the chlorine, you have a nitrogen group out there. So, chemically it makes sense that these would be -- they would be interacting with the receptor.

Comparing with the literature reference, there were these 46 chemicals against those that are tested in lots of labs, lots of literature, reference 93 percent or better, balanced accuracy. The ones that we missed were either classified as weak or very weak. Probably our active at near or above the concentration where we test. We simply couldn't see the activity.

Then comparing with the Tier 1 binding assay, we predicted the majority of the AR binding -- let's see, what did I see? We predicted a majority,

but not a lot of them. A lot of the missed -- in fact, that's actually not true. This was a mistype. We actually missed most of the actives, but most of the actives in this Tier 1 test were active at concentrations well above where we could test. There is at least an explanation for why we didn't see those. We had a good agreement with the negative results, except we had this cross-reactivity.

We have to talk about limitations. We don't have any metabolism in here, but there is this big ongoing challenge where EPA and NTP and NCATS are funding a bunch of groups to try to retrofit assays with metabolism. We're actually doing some of that work in-house. None of it is published yet, but it's -- in principle, we could take all of these 11 assays and we could retrofit them and actually run them with and without metabolism.

For estrogen, we actually took the -- sorry, I'm tired. We took the model data and we had a bunch of groups around the world build QSAR models, who could then go and predict what was going on. So what we've done relative to metabolism is we took those QSAR models, which are pretty accurate, and we took models that predict what the metabolites are, and

we're able to predict that the metabolites look estrogenic, the parents don't. So we took a set of reference chemicals where we know the parent gets metabolized to something which is more estrogenic.

Using this QSAR approach, we were able to replicate that trend. So it looks like maybe you have to do the experiments to predict what's going on with metabolism. Or at least for bioactivation to being something estrogenic, you can do a reasonable job using QSAR models.

We've talked a lot about DMSO solubility, so we've only done DMSO soluble to date, but we're working on water solubility.

Finally, in summary, we believe the model could today, with no question, be used to prioritize chemicals for further analysis using some combination of the AUC and the confidence score. We have to decide, okay, how do you put those two together? We don't have a final answer to that. This is one tool in this pivot using the high-throughput computational methods.

We're continuing to improve. For instance, we are running -- adding to the 11 assays, a cell proliferation assay, an androgen receptor

proliferation assay, which we had in the estrogen model. Hopefully, in the next few months, we'll have all the data there and we'll improve the model there.

We actually have more than the 1,800 chemicals in-house, so we can provide all of this data for a bigger set of chemicals. We believe -- this is the charge question -- that there is good enough evidence here that the current panel could be an alternative to the current test Tier 1 AR binding assay. A lot of people -- this is a subset of all the people who have worked on this model and the data and so on.

Any questions in that section?

DR. KRISTI PULLEN FEDINICK: If no one else has any, then I'll certainly -- thank you for answering all of these questions.

One of the questions I have, too, is about just the chemical universe. I'm not a computational biologist, it would be interesting to get you guys' thoughts on this. I did some simple Jarvis-Patrick clustering to kind of look at the EDSP universe as a whole, all 10,000 chemicals. For about 6,500 of them that I could do the clustering, there were about -- let's say 3,000 clusters that came up.

So the reference chemicals that you guys used only covered 36 of those -- were only in 36 clusters.

DR. RICHARD JUDSON: Sorry. How many clusters?

DR. KRISTI PULLEN FEDINICK: There were 3,000 total for all 10 -- about, again, 7,000 that I could get the structures for. Some of them are anise oil and you can't really get a structure for that. It's hard. If you look at the standards that you used, they were only in 36 of those 3,000 clusters.

So, if we really want to be able to look at whether or not these tests are reliable across -- again, we're not just looking at a small subset of chemicals. We're really looking across 10,000 potentially. Is there a way to really try to understand if that chemical universe has been -- or the chemical standards that you've used in those 36 clusters that are being generated, are the only ones that you would expect to be androgenic, or anti-androgenic?

I guess the question is, it doesn't seem to me as though 36 clusters is enough to be able to tell whether or not the chemical universe has been

expanded enough to apply that to 10,000; but are there scientific reasons or other reasons why that would not be the case?

DR. RICHARD JUDSON: There are sort of two ways to address that. One is, are our reference chemicals representative enough of the whole universe? That's really the bottom line. And, no. The reason is that just nobody has randomly tested lots of chemicals. So coming up with a broad set of reference chemicals is really, really hard. So the answer to that is no.

One question that -- here was a disappointment to me as a scientist. We ran the AR model on 1,800 chemicals, including many which didn't look at all like the reference chemicals. And the same thing for AR. We were going to discover something really new and cool, some estrogens and androgens nobody knew about, and we were going to be in *Science and New York Times* and all that. And there was nothing.

It turns out that the chemical classes, the typical chemical classes, if you go much outside of those, there just isn't much activity. That's one piece. And the 1,800 -- I don't know if you have done

this, but overlay the 1,800 on the 10,000, that's a much more representative set. I would bet -- but until we test them all, I don't know -- that that's not -- there aren't all these lurking pockets of androgens or estrogens out there.

The other thing we did -- which probably doesn't prove anything -- but we had these international groups build lots and lots of QSAR models based on the 1,800 chemicals. We actually pulled some literature data on another six or seven thousand chemicals. Again, we didn't see other structural classes that popped up that were consistently estrogenic or androgenic. So that's anecdotal information that we're probably doing okay. Our assay set is probably good enough to test the 10,000; but until you test the 10,000, you won't know.

DR. KRISTAN MARKEY: Not to preview too much, but exactly this question is coming up -- is it fair to advertise? I'm just going to advertise at this point that it is a topic for a future SAP meeting, to look across many of these clusters that we are also recognizing within there, and making sure that we have adequate space and coverage across those domains to confidently predict whether or not ToxCast

itself covers it, or we need other types of assays to cover those clusters that we see forming within that universe. That number 10,000 may or may not stay the same.

DR. KRISTI PULLEN FEDINICK: Can I ask a quick follow-up? Would the current Tier 1 tests -- again, if what we're thinking about is replacement here, would the current Tier 1 tests have the same limitations in terms of the ability to say, look across those different chemical classes potentially? Some of the physical chemical differences could result in volatility or other just characteristics of a chemical that make it reside outside of the ToxCast testing platform.

How do we -- and maybe you can't say this, right? And this is something that we have to think about as an SAP. But how do you think the current tests would be able to look across this chemical space in a way that might be different than ToxCast?

DR. RICHARD JUDSON: Again, my hypothesis is the more assays, the better. There are going to be a bunch of chemicals which goof up that binding assay, right? Without some kind of backup,

you will never know whether that's false positives -- whether those positives are false positives. I can't think of a way to -- you could also have false negatives for reasons I couldn't come up with right off, but you could potentially have false negatives there with the current binding assay.

There are two reasons why we're doing this. One is we're taking one assay, an in vitro assay, and replacing with 11. Maybe we can come up with a subset. This is a question that may come up later. You don't need all 11, but – so we're replacing one with many. But the one is sort of, from my understanding, kind of a hand crafted. It's not rolling off a log easy.

Whereas, with these high-throughput ones, you can -- at least if you're willing to run a lot of chemicals at once, you can take 1,800 or 2,000 or 3,000 chemicals and in a few weeks, you just get the answer for all of them. It's a practical -- with a testing program like EDSP, it's a practical problem how you actually do that though.

DR. REBECCA CLEWELL: I'm wondering if I can ask a question about the charge question. Can I do that now?

DR. JAMES MCMANAMAN: No. Soon. You can address it when it comes time for the charge question.

DR. REBECCA CLEWELL: I think it could go to some of these concerns though.

DR. JAMES MCMANAMAN: That's appropriate, if it's related to the charge question, okay?

DR. REBECCA CLEWELL: I won't ask. I want to say something else.

DR. JAMES MCMANAMAN: Okay.

DR. REBECCA CLEWELL: There has been a lot discussion about the technical limitations of these assay. I think -- sorry, I lost my speaker. One of the things -- I don't know if a comment is appropriate. Is a comment appropriate more than a question?

DR. JAMES MCMANAMAN: Just questions.

DR. REBECCA CLEWELL: Oh, man, you guys are such sticklers.

DR. JAMES MCMANAMAN: You can make the comment during the charge question.

DR. REBECCA CLEWELL: I'll wait for the charge question.

DR. JAMES MCMANAMAN: Thank you. If there are no other clarification questions, then we can move onto the next presentation.

It looks like a short presentation, so I think we can do that before lunch. Let your stomachs growl for a little while and we can get this done.

DR. RONNIE JOE BEVER: The androgen pathway model discussion.

DR. JAMES MCMANAMAN: This is Dr. Bever.

DR. RONNIE JOE BEVER: Yes, excuse me. Joe Bever. My discussion is going to be more in the frame of the regulatory. I thank Dr. Judson for presenting the science. He presented a lot of details for the androgen receptor model. Now I'm just going to describe how we're going to use this model, and why we feel like it's ready for that use.

This is, as Dr. Judson has pointed out, we intend to use it for prioritization. It's already been shown to be useful in providing some endocrine bioactivity data, which is one of the things that we're looking for in prioritization; the other being exposure. We also intend to use it as an alternative

for the low-throughput androgen receptor binding assay.

The context of use remains the same. We're looking at the androgen receptor binding pathway. With the pathway model, we're not simply looking at binding, however. We feel like the androgen receptor pathway model offers some advantages.

First of all, it's much more robust and informative. As Dr. Judson has pointed out several times, we have 11 assays rather than a single assay. It's useful for rapid prioritization. It reduces animal use. All of these assays are basically cell-free or exist with an established cell line. If we are using animal products -- and we do in one.

One of the assays is basically an upscaled, low-throughput assay where rat prostate cytosol is used in the low-throughput assay. So it would be used in the high-throughput, also, but at a greatly reduced volume.

Once again, animal use is reduced. It's high-throughput, so it's going to have the potential of saving money and definitely time, and also other resources. This pathway model is nice in

that it can show how the chemical is interacting, basically agonist, and antagonist. It gives you some idea about interference, and it gives you an idea about potency.

I'm presenting this charge question not to discuss it -- deliberation is not now -- whether it's just to remind you of the question and present the Agency's viewpoint on it. We have already delivered this androgen receptor pathway model to the SAP for deliberation in 2014. Now we're saying that if we met these recommendations of that SAP, then it should be suitable to act as an alternative for the low-throughput androgen receptor binding assay.

I promised you a more exhaustive list of what the December 2014 SAP had to say about our first generation androgen receptor model. Here it is. These are the major points from the minutes.

They said to evaluate cytotoxicity, so we ran concurrent cytotoxicity assays for the antagonist assays. As Dr. Judson has explained, the cytotoxicity assays are important to reduce any confounding due to cytotoxicity in the antagonism.

Now, these cytotoxicities, we can calculate the z-score. He explained how that is done.

And that is a measure of cell stress, which is useful for, as I say, removing any confounding factor here.

They asked us to expand the reference chemical AUC value range. Well, through our systematic literature review, we were able to almost triple the number of reference chemicals. We have the negatives, agonist and antagonist. We also have various potency. Thus, we fulfilled this recommendation of increasing the AUC range.

They asked us to optimize the assessment of activities. Once again, I'd like to point out that it has 95 percent accuracy, balanced accuracy. I'll get more into that later on -- especially with this reproducibility and transparency -- when I talk about fitness for purpose. But I will say right here, that yes, we've addressed these issues also.

They asked us to build on the assay battery, and we have. Previously we had nine assays. We added another couple, so now we have 11 assays in the battery.

The final point, as I previously mentioned, which is develop androgen receptor related assays that do not follow the classical genomic

nuclear receptor pathway. We haven't done, and the current low-throughput assay doesn't do it either. That does not preclude us from using it as an alternative to that low-throughput assay. It is a good idea and we may actually address that in the future.

Besides meeting the recommendations of the SAP panel, we feel like we also meet the standard validation principles. The validation principles of relevance, fit for purpose, reliability, and certainty and transparency. It fits the same context of use. It's looking at the androgen receptor binding and that biological pathway; but here, as far as relevance, we're looking at the same mechanistic and biological relevance of the original validated androgen receptor assay. Once again, it's basically looking at the same thing, which is androgen receptor binding. It's less or more.

Reliability. We have 11 orthogonal assays. These orthogonal assays include assays on not only receptor binding, but cofactor recruitment, RNA transcription, and protein production. So it's more robust. We're looking at different parts of the pathway. It's more robust and we have different

technical aspects to each of the assays. For instance, once something that might confound one assay, may not confound the other assay because of, say, the detection mechanism.

For reliability, we not only have all of this diversity, but we also have -- in multiple assays -- but we also incorporated, as was mentioned, concurrent cytotoxicity assays to remove a confounding factor. We also incorporated -- and this was really important in the Tox21 antagonist luciferase assay. We tested at two concentrations and we noted any shift in the curve, which would be telling of true antagonism versus a false antagonistic reaction.

So, for reliability, all of these different factors make us confident of the output of the model. The model is basically using almost like a way of evidence approach, as Dr. Judson mentioned, through these multiple assays and diverse technology.

We also measured uncertainty, which Dr. Judson showed. And this was characterized with a bootstrapping procedure and the amount of uncertainty was not really dependent on the model, it was dependent on the chemical. You could see the

different chemicals and how they had different confidence intervals.

Lastly, transparency. We have a very detailed explanation of the androgen receptor model in the white paper and supplemental files and linked sources. We hope that this is adequate for everyone to understand the methodology we used, as well as being able to reproduce our statistical analysis and modeling.

We feel like this pretty much met the validation principles. Including fit for purpose. We had a greater than 95 percent balanced accuracy. This is assuming the ambiguous findings are positive, which is health protective.

This is using 46 reference chemicals found through systematic literature review. Dr. Judson has explained, when comparing it to the low-throughput assay, that generally there is a great correlation, a great agreement. And when there is not, we were able to give a good explanation for most of those cases why there was disagreement.

In conclusion, the Agency feels like that we've implemented all the pertinent SAP suggestions from December 2014 in regards to the

androgen receptor pathway model. We also feel like the basic validation principles have been met. Thank you.

DR. JAMES MCMANAMAN: Any last questions for Dr. Bever? Dr. Clewell?

DR. REBECCA CLEWELL: Now I have my charge question question. I think maybe just to make sure we're all on the same page. The charge question this week is specifically about the replacement of the low-throughput androgen receptor binding assay, which currently within the EDSP testing schema goes together -- is used together with some in vivo studies to determine androgenicity. Right now we're only discussing the replacement of the low-throughput androgen receptor binding assay, right?

DR. RONNIE JOE BEVER: Yes, but as I said about presenting the charge questions -- which I'll do again with the steroidogenesis. My purpose there was basically to show the Agency's viewpoint on it. The deliberation on the charge questions will occur later. But, yes, we're simply looking to replace the low-throughput androgen receptor binding assay. Or not replace actually, an alternative for.

DR. REBECCA CLEWELL: And it would be then for future chemicals, you could either use the data that's already been collected through the ToxCast effort, or you would collect data in the 11 assays and develop a new AUC score, and that would be your decision making number? Is that --

DR. RONNIE JOE BEVER: As you recall from my background, there is weight of evidence analysis for the Tier 1 battery. This would be an alternative for a Tier 1 test. Yes, the AUC is important, but that's certainly not all we consider before we say something is bioactive and requires the Tier 2 testing.

DR. REBECCA CLEWELL: That makes sense. I was actually just wondering about the collection of the data. Like the recommendation would be that we would -- instead of doing the low-throughput AR binding assay -- by we, I mean someone, not me -- would collect data in 11 ToxCast assays, develop an AUC score and that would be that sort of input for androgen receptor binding? Yes? Okay. Thank you.

DR. JAMES MCMANAMAN: You're welcome. Other questions? If not, then I think we'll break for lunch. It's 12:20; we'll be back at 1:20.

[LUNCH BREAK]

DR. JAMES MCMANAMAN: It's time to get started again. Remember, any ex parte conversation should be read into the record, unless it has to do with dinner or something like that.

Dr. Paul Friedman is up next. So, if you're ready. It looks like we have a few stragglers, but I think we've got a quorum of permanent panel members.

DR. KATIE PAUL FRIEDMAN: Thank you for the opportunity to present today. I'm Katie Paul Friedman. To my right is Dr. Woody Setzer. We're both from the National Center for Computational Toxicology. I'd like to talk to you about the work that we've been doing to develop a high-throughput H295R assay, and then to statistically integrate the multidimensional readout that we get from that assay for application to prioritization.

Just a brief overview of my talk. I'll talk a little bit about objectives and the overall approach, the assay background and method. So, this will be a high level overview of how the high-throughput H295R adaptation works. This was published in 2016. Then the methods and results part

of this talk is split into two sections. First, we go through an evaluation of the high-throughput H295R assay comparing the OECD reference chemicals and their performance regarding the synthesis of estradiol and testosterone in the H295R system. Then the second part will be development of a quantitative prioritization metric for those high-throughput H295R assay data. And then a little bit of discussion and conclusions.

Objective one is really set up by this initial challenge that within the EDSP Tier 1 battery there already exists a low-throughput H295R steroidogenesis assay. There are both EPA and OECD test guidelines for this steroidogenesis assay. The assay is really employed to look at potential perturbation of estradiol and testosterone.

Our initial objective was to adapt that assay to a high-throughput format to increase resource efficiency and speed, to address the questions that have been discussed at length this morning about the too many chemicals, too little time problem.

Objective two, within this high-throughput version of the assay, we actually were able to measure 11 steroid hormones synthesized in the

cell. We wanted to utilize all of that information. So our second objective was to develop a summary measure that integrates the multidimensional data to quantify pathway perturbation, and then hopefully indicate the relative priority for further screening or evaluation of chemicals based on their potential effects on steroidogenesis in this model.

This is an outline of our overall approach, but also to highlight for you the publications that are really included in the white paper that you have and in our talk today. The initial high-throughput H295R assay and the stage screening approach that we employed for resource efficiency, was published in 2016 in *Toxicological Sciences* by Agnes Karmaus and colleagues. The paper is referenced here. I'll go over this a little bit in the assay background and methods.

The rest of the talk today is really described in a paper that was actually just accepted yesterday at *Toxicological Sciences*. That will be Haggard, et al. 2017. That paper goes through first the evaluation of the high-throughput H295R assay via a comparison to the OECD inter-laboratory validation results that were published in 2011. Those data

basically are analyzed per a very similar protocol to the OECD test guidelines, so that we can perform that more apples-to-apples comparison with reference chemicals, and then evaluate the concordance of estradiol and testosterone responses.

The second part of that paper and the second part of this talk, is the development of a prioritization metric. What we have done is statistically compressed the data for an 11 steroid hormone panel, measured in the H295R cells and then evaluated that prioritization metric.

Any questions before I go onto assay background? Okay.

This section, as I mentioned before, is really about just describing that Karmaus, et al. 2016 methodology that we implemented to screen now 2,012 chemicals through the high-throughput H295R assay.

First, for those of you who aren't familiar -- and I imagine everyone on the panel is familiar to some extent with the importance of steroidogenesis -- but this is essentially the process by which cholesterol is converted to steroid hormones. This is really important physiologically for sexual

differentiation, development, reproduction, but also basal metabolism.

There are four major classes of steroid hormones synthesized largely in separate tissues in vivo: progestogens, corticosteroids, androgens, and estrogens. The disruption of any of these can result in the development of a wide range of disorders including congenital adrenal hyperplasia, or effects on fertility, or even hypertension and metabolic functions. This is very completely reviewed elsewhere.

Steroidogenesis is the H295R model includes all four classes of steroid hormones, so this is a really unique model that you can see is used to our toxicological advantage. Here is just a representation of the model in colored quadrants here. Green are progestogens, blue are androgens, yellow are corticosteroids, and red are estrogens. All four of these classes are generated within the cell and can be measured in the medium in this assay.

The steroid hormones written in black text are those included in the pathway that we've included in the high-throughput version of the assay. I've highlighted in white the two hormones that are

typically measured in the low-throughput guideline version of the assay, which are just testosterone and estradiol. We measure those as well in the high-throughput version. The key difference here is that in the high-throughput version, more steroid hormones are measured.

Our first implementation of this assay maximized screening resource efficiency. To do this, we actually -- very similar to the NovaScreen assay technology that was described previously -- we first performed a single concentration screen at a high concentration. To do this, we determined a maximum testable concentration for each chemical. We defined that maximum testable concentration, or MTC, as the concentration that maintained a minimum of 70 percent cell viability. Then we screened that concentration for effects on any of the steroid hormones. From that set then we advanced -- now 656 out of over 2,000 chemicals have been screened in multi-concentration response.

I'd like to walk you through the graphs here just a little bit on this slide. The first graph on the left is a demonstration of the concentration ranges that we observed for the maximum tested

concentration. On the X-axis is just number of unique chemicals. And you can see here for the vast majority of chemicals, the maximum tested concentration was somewhere between 10 to 100 micromolar. But there are some chemicals for which we had to revise down the maximum tested concentration.

The graph on the right is the number of unique chemicals versus the sum of positive hits in single concentration screening. One of the criteria that we used to advance chemicals beyond single concentration screening was to look at the sum of steroid hormones that were affected by that particular chemical. Most of the chemicals advance affected three or four, or more, steroid hormones in the set. There are some that we advanced that were negatives or references, but a good number of them perturbed three to four at least of the hormones in the set.

This allowed us, as you can see, to reduce the resources used to screen because approximately half of the library we screened didn't affect any steroid hormones in the set at that max tested concentration.

This is just an overview of the method itself, very briefly. The cells are plated overnight

at 50 percent confluency and just allowed to incubate. Then a key difference in our assay methodology from the guideline version, if you're familiar, is that we perform a pre-stimulation with forskolin, abbreviated here as FSK. For 48 hours post-plating, there is a 10 micromolar forskolin stimulation. This basically upregulates the entire steroidogenesis pathway in the cell, to try to get an increased signal since we've miniaturized the assay.

Then there is a washout and a chemical exposure for 48 hours, and typically our max tested concentration approaching 100 micromolar. And then at the end of that 48-hour period, the cells are evaluated for cell viability using an MTT assay. The medium is sent to a contract lab called OpAns where HPLC tandem mass spec is used to quantify actually 13 hormones. I'll talk a little bit in a few more slides about the 13, or is it 11, how many hormones? The quick answer is we tried to measure 13, but we can only use the data for 11. I'll talk more about that.

I wanted to point out for you -- in the white paper there are some very detailed tables and text about the methodological differences between the guideline version and the high-throughput version.

What I've tried to pull out for you here are the primary differences between these methods. I've sorted these differences kind of by their stage in the approach and which aspect they relate to. Then here I have the OEC test guideline 456, and then the actual implementation that we used in our screening.

The key difference is that forskolin pre-stimulation that you don't typically see in the guideline version of this assay, in addition to the fact that we've miniaturized the assay to a 96 well plate version. Typically, contract labs are running this in much lower density, like a 24 well or maybe a 48 well plate. So we've really increased the throughput on that.

There are a few other differences here like replicates. Again, to increase our efficiency, we have fewer biological replicates, we have fewer technical replicates. I have some slides at the end today to look at reproducibility, and actually we are capturing, I think, the variability that we need to capture in order to reproduce the results.

Then acceptable cell viability. Some have pointed out this minor difference here where in the guideline it says that you need to have your

minimum cell viability be 80 percent. We've allowed for 70 percent.

The brief answer for that -- and if there are questions, I can go into more depth -- is that based on the median absolute deviation around the baseline of the MTT assay, it would be difficult to discern effect of less than 30 percent. Considering that there is noise around the baseline and we've screened over 2,000 chemicals with the MTT assay, we can really define what baseline noise looks like, and so that's why that 30 percent threshold was selected.

Are there any questions? I see some nodding around the -- all right.

Finally, when you're considering development of a high-throughput assay, you really want to consider quality metrics that would indicate to you that that assay is robust enough to use in a screening environment to know that you could actually distinguish signal from noise. And that you can get a sufficient effect size in the correct direction. That's what this slide aims to explain.

The table presented here is actually a reproduction of Table 2 from Karmaus, et al. 2016. On the left are the steroid hormones that were reported

in that publication. And then I have the Z-prime or Z-prime factor, and the strictly standardized mean difference shown for two different reference chemicals.

Z-prime, as you know, is a measure of sufficient signal to background distance in low enough variability that you could actually distinguish a reference chemical from simple noise. The strictly standardized mean difference is a measure of effect size and directionality. You can see here for forskolin, which is the stimulatory reference chemical, and prochloraz, which is the more inhibitory reference chemical, we have Z-prime factors of generally over 0.5. In some cases, approaching 0.8 or greater. Any Z-prime factor greater than 0.5 is an assay that you would probably be able to run in a high-throughput screening environment and be able to distinguish signal. So we think the performance there was quite good based on these reference chemicals.

For the strictly standardized mean difference, or SSMD, the absolute value of that was greater than seven a majority of the time. We think this demonstrated robust effect size and directionality. You can see for prochloraz, this is

an inhibitory chemical for most of the pathway, and we see a generally negative value.

You can see an exception here is progesterone, where actually prochloraz appears to stimulate it. But for many of the hormones it's inhibitory, and we see the correct direction of effect. So, taken together, these assay quality metrics signal to us that this would be an appropriate assay in a screening environment.

Are there any questions on that section?

DR. SCOTT BELCHER: I had a question about the forskolin stimulation. Is this stimulation above your pre-treatment level? Is this a second stimulation or is it that single stimulation? I don't understand that.

DR. KATIE PAUL FRIEDMAN: It's a single stimulation event right after plating for 48 hours.

DR. KRISTI PULLEN FEDINICK: I had a question about the pre-screening process. In the -- is it Karmaus paper? Is that how you pronounce that last name? In the Karmaus paper they found that when they pulled from the zeros, right -- so, the ones that had no reaction or no effect in the initial high dose

prescreen -- that 53 percent of those had an effect when they went back and rescreened them. So, they just pulled those out.

It makes me wonder about the prescreening process and if you're -- if you have 86 percent accuracy, let's say with that, but you're missing 50 percent or more of your chemicals that only had a single dose actually did have activity, is this prescreening throwing out things that could ultimately have an effect later on?

DR. KATIE PAUL FRIEDMAN: To clarify a little bit, I think I know the table that you're talking about from that paper. It's a supplemental table and it looks at the recall sensitivity. It looked at if you had an effect on four or greater hormones, the recall sensitivity, i.e. your ability to get the same effect a second time, was 86 percent. So, 86 percent of the time you got the same effect. Whether it was positive or negative, there was agreement, there was concordance.

But then when you dropped to an effect on three hormones or fewer, the recall sensitivity dropped to 53 percent, meaning that it became harder to replicate the same effect. Basically the way I

interpret that is that you see for chemicals that act weakly in the pathway, you can see some borderline responses and borderline responses are always hard to replicate.

From a biological perspective -- that's kind of the statistically reasoning, right? But from a biological perspective, let's say you inhibited aromatase. You inhibited SIP-19A1. You would imagine actually -- in theory, you would impact four hormones because you would impact the production of estrone and estradiol, and also you would impact the relative concentrations of these androgens. From a biological perspective, you might hypothesize that that might be a limit that would be of biological interest.

Later on in development of the prioritization metric, we have a Venn diagram that looks at the number of chemicals that affected each class of steroid hormones. Actually, for most of the chemicals that we screened in concentration response, they affected more than one class.

Yes, in theory, you could miss something, but it's likely that it could be borderline -- a borderline response that is hard to replicate. Of course, given more resource, you could go back and

you could probably find some chemicals that you could rescreen. Everything is always about resource efficiency in screening, and so of course, you could consider ways to revise that and try to pull more lists.

DR. KRISTI PULLEN FEDINICK: Just to clarify. In the paper they said that 64 of the additional 121 chemical samples that didn't meet the concentration response selection criteria altered at least one hormone. They took their initial 411 chemicals, or whatever it was, and then pulled an additional 121 chemicals out of those that didn't have a response and then ran those again and found that there was a response -- or 53 percent of those had a response. Did you do something similar?

DR. KATIE PAUL FRIEDMAN: It's recall sensitivity. It's not whether or not they had a response, it's whether or not the response was concordant between trials.

DR. KRISTI PULLEN FEDINICK: Well, it says the selection sensitivity.

DR. KATIE PAUL FRIEDMAN: Right, so it's the recall.

DR. KRISTI PULLEN FEDINICK: But then those -- let's say 64 out of 121 is a little bit more than 50 percent, right? Those had a response -- so, those chemicals wouldn't have been included in that initial screen because they would have failed the prescreen?

DR. KATIE PAUL FRIEDMAN: Right.

DR. KRISTI PULLEN FEDINICK: My question is, did you do something similar for this where you -- if you took the 626, or however many chemicals you had -- did you pull also from the pool that failed your prescreen to go back and see whether or not you had responses? And these were randomly selected, those 121.

DR. KATIE PAUL FRIEDMAN: We've added some chemicals since the Karmaus, et al. paper. There are another 85 chemicals that we've screened in concentration response since then. So the numbers have slightly bumped up since that publication was released. Within the 656 are the 524 that were in Karmaus, et al., so they're the large kernel to what we're presenting here.

Basically the expansion since the Karmaus, et al. paper was to include chemicals of

interest, like phthalates, we included more of that class. We included chemicals that perturbed three hormones. In the Karmaus, et al. paper, they had a cap at four. That was really resource limited. So, we were able to reach back and grab those that perturbed three. So, as additional resource was put in, we modified to be able to test more.

DR. THOMAS ZOELLER: Because the product of one enzyme reaction becomes the substrate for another -- and you kind of touched on this with respect to aromatase. But when you look at the data, can you identify specific probable -- specific enzyme steps that are affected by looking at a reduction in one set of steroids and an increase in another as you were pointing out?

DR. KATIE PAUL FRIEDMAN: That's actually an excellent question and we have worked a little bit on that issue. The way I understand your question is, basically, can you pull out select patterns that suggest mode of action?

I think that's part of the goal here, eventually, would be to do that. But there were so many patterns that it became difficult to discern single enzymes that would be acting.

Some of the continuing work that's going in my group, and Woody Setzer's group, is to look at, could we build a kinetic model of the high-throughput H295R assay. If you're familiar with this area of toxicology, you know that there are several papers that have already examined the development of kinetic models and hypothesizes about mechanism of action within this cell line. They were optimized for a different version of the assay that didn't include pre-stimulation and was not miniaturized to 96 well. So, we've actually collected a little bit of time course information to see if we could build a kinetic model, but that right now is relegated to future interest and work.

The focus today, actually, when we had this set of hormone data was to think, okay, if we don't understand exactly all of these patterns and what each pattern means, can we develop a greater meaning from the whole set, in terms of which of these chemicals would be the highest priority to look at. So the reasonable approach seemed to be to look at considering the magnitude of effect on the whole pathway, and not excluding any quadrant and just looking at magnitude of perturbation.

DR. JAMES MCMANAMAN: I have a question about the MTT assay. Why did you choose that assay as an assay of death or of loss of cells? Because actually what it's based on is mitochondrial function, so you can have the same number of cells and decreased mitochondrial function will get a different MTT result. Since these steroid hormones are being synthesized, it requires mitochondria, it seems like it's a complicating --

DR. KATIE PAUL FRIEDMAN: Right. The question of a cell viability assay is a good one. The MTT is commonly used in contract labs that run this.

DR. JAMES MCMANAMAN: Yes, I know, but incorrectly.

DR. KATIE PAUL FRIEDMAN: But also you would want functional mitochondria, because if you had mitochondrial toxicity specifically, that would confound steroidogenesis. So I would imagine that that would be an appropriate cell viability assay.

DR. JAMES MCMANAMAN: Right, but I don't know whether it's cell viability or mitochondrial function. It seems to me to be important to know the difference whether it's one or the other, because it affects your interpretation.

DR. KATIE PAUL FRIEDMAN: Yes, that's an interesting point, and we could consider that further.

DR. REBECCA CLEWELL: Can you go to the slide that shows the staged screening? Slide 97? Do you have the same numbers as me? Thank you.

The plot on the right is used -- if I heard you correctly; and I may not have, so please correct me -- is to justify a minimum cutoff of three or more hormones to be considered for further screening. Is that right?

DR. KATIE PAUL FRIEDMAN: Yes, with some caveats. Most of the chemicals screened in concentration response come from where this blue box is outlining where those chemicals perturbed three or more steroid hormones. Some of the chemical screening concentration responses, as Dr. Pullen Fedinick already mentioned, were pulled from the sort of negative, the zero to two slot. Some were pulled from reference chemical lists, and some were added simply because they were chemical classes of interest. So this wasn't the only criteria, but it was the main criteria.

DR. REBECCA CLEWELL: I have some questions about that, because the idea that -- so, 656 chemicals -- or potentially somewhere around there -- have three or more. But at least 400 have one or two hormones that have changed, according to this plot here. Right?

DR. KATIE PAUL FRIEDMAN: Right, you're right.

DR. REBECCA CLEWELL: So, there's not that big of a difference to say the majority of the chemicals had three or more? I would say about 400 of them had between one and two. Am I reading this plot wrong?

DR. KATIE PAUL FRIEDMAN: I guess I'm confused by the question, so maybe I'll just rephrase. The reason that this was done was to maximize resource efficiency. So, you can see that there were a number of chemicals that hit zero steroid hormones, and then a number that hit one or two. So those were considered a lower priority for multi-concentration screening. So we proceeded in an iterative fashion, screening chemicals. Most of that set affected three or four or more steroid hormones.

DR. REBECCA CLEWELL: Did you guys look into whether you could classify -- I understand what you're saying about there being a sort of intuitive biological reason why you would see three or four changed at any one time, because this is all related. On the other hand, was there sort of an evaluation of whether these chemicals where you see one or two or more towards the terminal end of the pathway?

DR. KATIE PAUL FRIEDMAN: Right. So, would you see only an effect on, say, estrone and estradiol?

DR. REBECCA CLEWELL: Yes.

DR. KATIE PAUL FRIEDMAN: We could go back and -- I haven't asked that specific question. I have looked back through the list that we screened in single concentration to see how many affected only estrogen and androgen. That would essentially be a list that you could consider doing more follow-up screening on. But, as I mentioned before, there is a Venn diagram later in my presentation that shows that, really, a lot of the chemicals there were highly potent and acted within this pathway, hit multiple pathways. So, that wasn't really the norm for what we observed.

DR. REBECCA CLEWELL: I'll wait until I see the Venn diagram before I ask about that. The other sort of -- the question that sort of comes out of what you said, then, is so moving forward -- and I understand the need to use resources wisely. I don't have a lot of money myself, right?

But the question is, moving forward, if you were moving into application in an EDSP type program, would that be something you would recommend, is that we have a minimum cutoff of three hormones that are changing in the pathway before we consider moving forward? Is this high concentration testing, and then a cutoff of three or more hormones, would that be the sort of schema that you would recommend for an EDSP type situation?

DR. KATIE PAUL FRIEDMAN: I'm not sure of the answer to your question. I think that if you were using this in the EDSP realm, maybe you would have the resource to say, this is the chemical list of interest and so we'd like to have it completely screened in multi-concentration.

I think depending on the resources available, probably if you had a list of chemicals

that were of concern, you would want to do the multi-concentration screening.

Keep in mind, we were working with a library that was much bigger and subject to broader screening like Richard Judson mentioned. And we weren't working from a specific EDSP list. But if we were, you might implement the assay differently. You might just go straight to multi-concentration after doing a little testing to see what your max concentration could be.

DR. REBECCA CLEWELL: Thank you.

DR. JAMES MCMANAMAN: One more question?

DR. KRISTI PULLEN FEDINICK: Yes. Did you retest the chemicals that were run in the Karmaus paper, or did you only do 130 to 140 extra?

DR. KATIE PAUL FRIEDMAN: No, this is the same data, but as you'll see, we've analyzed it quite differently.

DR. KRISTI PULLEN FEDINICK: But you did run new tests on only a very small subset, so there is no way for us to, say, go back and kind of come up with biological replicates because you reran the same exact ones that they did?

DR. KATIE PAUL FRIEDMAN: Actually, later in the presentation I have a slide on chemicals. About 16 percent of the library was tested in more than one experimental block, and that's how we were able to examine reproducibility, because we do have biological replicates for those chemicals. So I have a slide directly toward your question at the end.

DR. KRISTI PULLEN FEDINICK: But you only added an additional 100 or so chemicals to this particular exercise? Or these were all from the Karmaus paper?

DR. KATIE PAUL FRIEDMAN: Since the Karmaus paper we've added a number of chemicals.

DR. JAMES MCMANAMAN: I guess we can move on.

DR. KATIE PAUL FRIEDMAN: Great. The next section of our talk is about comparing the results of reference chemicals used in the OECD inter-laboratory validation study, with the results of the high-throughput H295R assay.

This really aims to answer this question, which is, does the high-throughput H295R assay replicate the estradiol and testosterone results of the low-throughput assay. I think this is a fairly

salient question and it speaks to whether or not you could use this assay.

To do this, we performed a comparison to the reference chemicals used by the OECD and their published inter-laboratory validation study. Obviously as mentioned before, only two hormones were available for this comparison.

The high-throughput data were analyzed by a completely different method here than in the Karmaus paper. So I just want to make that distinction. The OECD test guidelines specifies a way to analyze these data. And so these data, instead of using the ToxCast data pipeline, as mentioned previously, were actually analyzed by ANOVA and post hoc Dunnett's procedure. The DMSO control data from the same plate were used for the sample comparison.

We used the same criteria for positive, as what was used in the Hecker, et al. paper, which was that two consecutive concentrations had to produce results that were significantly different from that DMSO control, or a positive at the max concentration that maintained at least 70 percent cell viability. We also applied a 1.5-fold change from DMSO control

threshold, which was applied to give context for very small changes in steroid hormone synthesis.

This comparison is essentially a binary comparison. We're comparing positive and negative between the reference chemicals and the high-throughput version of this. So, to do that, we've constructed some confusion matrices.

The OECD inter-laboratory validation results are from that Hecker, et al. paper, and were interpreted from Tables 3 and 4 in that published work. To give you an idea of how many chemicals were available for the comparison, the OECD inter-laboratory validation used 12 core reference chemicals. These were tested in five labs for that paper. And 10 of those 12 were screened in the high-throughput version of the assay.

Additionally, in the Hecker, et al. work, there were 16 so-called supplemental reference chemicals, and 15 of those were screened in the high-throughput version. The reason these are called supplemental, is because they were screened in only two test laboratories.

And this seems a little myopic, but these details are actually important because sometimes

the labs were not concordant. So, if you only have two labs, you're really left with an equivocal finding. And if you have five labs, but two or three don't find a lowest effect concentration, then you're again left with an equivocal finding. That's reviewed briefly here.

I think overall, the confusion matrices demonstrated good sensitivity and specificity, and a very good accuracy for the reference chemicals. You can see here at the top there is a table that lists the effect. I have divided this out by increased testosterone, testosterone up, and decreased testosterone, or down; and similarly, increased estradiol and decreased estradiol.

You can see that the accuracy all the way to the right, tends to be 0.8 to 0.95 in a best case scenario, depending on the effect type. In terms of the sensitivity, there are some case where we do extremely well, like testosterone up where they're all correct. Testosterone down, we had a more limited sensitivity; but, again, typically approaching pretty high numbers and a very good accuracy.

I'll put the accuracy in more context of the actual concordance between the labs themselves

than the OECD inter-lab validation to show you more graphically. These are the matrices that we developed. HT stands for the high-throughput data, and OECD obviously the Hecker, et al. data.

One commonality across all four matrices that you can see, is that there are a very limited number of true positives in relationship to the true negatives. That's a little bit of an unbalanced set for each effect type.

As I mentioned, putting this into a little bit of context, when you have those accuracy values that approach from 0.8 to 0.95, what does that really mean? What is an accuracy value that's really good enough? One way to ask the question would be to look at the actual agreement or concordance among the labs in the inter-lab validation itself.

So, for any effect on testosterone, the average concordance among labs was 0.88, 0.91, and 0.90, for the core reference chemicals only, the supplemental reference chemicals only, and then the entire set together respectively. So, around 0.9.

And then for any effect on estrogen, similarly, the average concordance among labs was 0.95, 0.84, and 0.89, for the core reference

chemicals, supplemental reference chemicals, and the entire set together.

Again, you see this kind of threshold at around 0.9, which actually we come very close to here with accuracies ranging from 0.8 to 0.95. We're probably reaching the limit of the ability to actually predict the OECD data with the high-throughput data, just based on the fact that not all five labs agreed each time, or were able to find a LOEC. This doesn't even speak to, kind of, the potency range that was found for each chemical. This was just a binarization of the data of positive and negative.

Are there any questions on that section before I move on?

DR. KRISTI PULLEN FEDINICK: I have tons of questions. For the testosterone down in particular, so you missed about 35 percent of the true positives, right? And this is your revised sensitivity. So, in the unrevised, you got rid of the nonoxynol-9, which made sense because of the chemical. But I wasn't clear as to why you got rid of -- what was the other chemical? Letrozole.

DR. KATIE PAUL FRIEDMAN: Letrozole? Yes, so letrozole was only removed from one of the

testosterone matrices because the only effect was found at 100 micromolar, and our MTC was like 1.4 micromolar. So, it was orders of magnitude greater than what we were able to test. So that's why I excluded it.

And then nonoxynol-9, as you mentioned, for simplicity here in the presentation, I didn't go into it, but you're right. These are the revised matrices excluding nonoxynol-9 due to uncertainty in what the chemical tested in the OECD inter-laboratory validation actually was, and what molecular weight it would have corresponded to, and what structure. And letrozole was removed due to very large differences in the ability to test a max concentration.

DR. KRISTI PULLEN FEDINICK: For the T down, essentially for the unrevised, it was a coin toss as to whether or not you got the true positives, right? If it was about 50 percent?

DR. KATIE PAUL FRIEDMAN: Actually, I have another slide deck with those numbers in them. Sorry, please hold for the computer. These are the confusion matrices prior to the removal of nonoxynol-9 and letrozole. You can see what Dr. Pullen Fedinick is talking about.

If you look at the testosterone down here, the sensitivity was 0.55 because we removed those two chemicals that had clear issues in our ability to screen them. Nonoxynol-9, because we aren't sure what the OECD screened. And letrozole, because there was a very large difference between our ability to test the concentration that they tested. It would have been cytotoxic at that concentration for us. We thought better to leave it as equivocal than to exclude it.

I think really the revised numbers are more reflective and more indicative and that's why I chose to include those in the presentation. But for the sake of transparency, I wanted to show within the paper that we did it two ways.

DR. KRISTI PULLEN FEDINICK: Can I just ask one follow-up? Just really quickly, for the accuracies that you showed -- I forget what slide that was; I don't have it up right now -- for the inter-laboratory validation. Do you know whether or not the sensitivity and specificity, what the values were for that? This is slide 106. If you just have your -- this is the average concordance -- maybe I'm missing this.

Do you have a sense of what the sensitivity was for the other tests? If we can compare apples to apples rather than just the balanced accuracy. If I'm more concerned with sensitivity, for example, as the EDSTAC recommended that the EDSP do, then how can we tell whether or not that balanced accuracy is really being made up by the increase in specificity, or if it's how our sensitivity is doing in those other valleys?

DR. KATIE PAUL FRIEDMAN: I'm not sure I follow the question in there. There is a slight reduced sensitivity there for testosterone down. I guess I would also say that this perspective, while helpful in looking at fit for purpose validation, is only one dimension. If you're simply binarizing data into positive and negative based on an ANOVA and then whether or not you caused a 1.5-fold change, that's one way to look at the data that can be informative.

I would argue that actually the second half of our presentation is more quantitative and would relatively rank chemicals based on a metric that reflects the effects size and, to some degree, also the potency, would really be a more useful metric for use in the program. So I think I might separate in my

mind this approach as one approach that's commonly employed for fit for purpose validation versus screening the EDSP library and how we might prioritize them.

You might find a chemical that should have been a positive that was missed in high-throughput; but perhaps it's a pathway positive, so it affected other steroid hormones. Actually, we see that. One of the chemicals missed for an effect on estradiol is mifepristone. But it's an extremely strong agent against progestogens, so in the pathway approach we catch it and then we can relatively rank it using a quantitative value.

I think this is one way of comparing to the reference chemical performance, but I think a quantitative ranking that includes the whole pathway is another way that bolsters confidence in its use programmatically.

DR. JAMES MCMANAMAN: This slide will be put into the docket, so we'll have that information available. Can we move on?

DR. KATIE PAUL FRIEDMAN: Without further ado, we'd like to talk a little bit about solving that 11-dimensional problem. We've looked at

just estradiol and testosterone. We showed good accuracy in predicting the OECD reference chemical effects with the same chemical screen and high-throughput. But now we'd really like to expand beyond those two hormones to looking at the 11 steroid hormones that we could measure in this assay.

This slide speaks to a little bit of the numbers. This is a comment that I've received repeatedly, that we have a lot of numbers and we toss them around and we change. I sympathize. There were 13 hormones that were measured in the high-throughput version of the assay. However, pregnenolone and DHEA were very often at concentrations that were below the lower limit of quantitation; in fact, 53 percent and almost 70 percent of all the measurements. And because of that, these two hormones were excluded. They just weren't present enough of the time to consider the data reliable. So, because of that, we used 11 hormones in this analysis.

To confuse it further -- I didn't put it on the slide -- there are 10 hormones in the Karmaus, et al. paper. We've done a reanalysis of the data where we're using the DMSO control and analyzing

the data completely differently. So, we had different criteria for inclusion.

Here is an example of the 11-dimensional results for prochloraz, plotted in a very traditional way that we will all recognize. On the Y-axis is the measured analyte in micromolar. And then we have along the X-axis the concentration of the chemical, prochloraz. Then we have 11 different plots for 11 different steroid hormones.

You can see, as discussed previously, typically it's an inhibitory action of prochloraz, except for a few of the hormones like progesterone. There is a stimulation of the synthesis. But particularly for the androgens, estrogens, and corticosteroids, there is an inhibitory action.

But, ideally, you can't really look at 11 plots all the time for 10,000 chemicals that you might want to rank. It's not very practical, and how would you decide? So, you need a mathematical way, a statistical way of compressing this into a value. That was our goal.

What we can learn from these other steroid hormones, it has been coming out a little bit in our conversation, but there are other biological

advantages to including these data. For instance, additional evidence for disruption of estrogen or androgen synthesis. If you were to upregulate aromatase like atrazine does, then you would see an effect on estrone and estradiol. And seeing that concordance across two analytes might give you greater confidence that that's what happened.

As discussed with Dr. Zoeller's question, you could possibly see punitive mechanisms of steroidogenesis disruption. And in some cases this is more clear than others. Prochloraz it's very clear, and for the conazole fungicides it's very clear that across the board there is a very strong inhibitory action. You could cluster chemicals together and say this chemical kind of behaves like these other chemicals, and have some learning there.

Finally, you might also learn something about effects on other specific steroid hormone classes, namely the corticosteroids and progestogens. Within this model, although it's artificial, you can imagine that given enough concentration and time, an effect upstream on progestogens would propagate to the biosynthesis to the downstream androgens and estrogens.

This is the promised Venn diagram. I wanted to show that most of the screened library actually affected multiple steroid hormone classes. This Venn diagram includes the chemicals screened in multi-concentration. Actually, if you have your calculator, this will only add up to 628 chemical samples. There were 653 chemicals that we were able to include in this analysis out of the 656 that affected at least one steroid hormone. So, those that affected zero are not in the Venn diagram.

What you can see here is that there is one chemical that affects androgens alone. There are eight chemicals that affect estrogens alone. Only one chemical that affected only those two steroid hormone classes.

Much more compelling is the 307 number in the middle, which suggests that a very large percentage of the chemicals actually affected all four steroid hormone classes, which you would expect given the criteria that we used in the staged implementation of this assay. Very few of the chemicals affected only progestogens or only corticosteroids, or even only the union of those two.

The metric we chose to compress this 11-dimensional data into one dimension is called the Mahalanobis distance. And the Mahalanobis distance essentially adjusts for the variance and covariance among the steroid hormones in the set prior to calculation of the Euclidean distance. Really the purpose of this is so that you can understand the effect size and have it be truly representative, rather than difficult to discern based on the correlation of the residuals on these measures.

You can imagine that the residuals on these measurements might be correlated for a couple of different reasons you might hypothesize. The hormones were measured from the same experimental well, from the same medium. And the synthesis of these steroid hormones is obviously interdependent in that pathway.

The Mahalanobis distance then adjusts the distances or effect sizes, accounting for knowledge of that interrelatedness of the steroid hormone measurements, without biological knowledge, just knowledge of the correlation of the residuals on the measures.

I wanted to provide some support that this is actually true. That the residuals for the

steroid hormone measures are correlated. And this heat map does that. I'm going to walk you through it a little bit.

The top half above this diagonal is actually the same as the bottom half. You can only focus on one, if you like. Basically, the deeper the blue color, the greater the positive correlation between the residuals on the measures of the hormones. For instance, estrone and estradiol had highly correlated residuals with a Pearson's R of 0.75. Androstenedione and testosterone had a very high correlation, as well as cortisol and 11-deoxycortisol. You can see some examples where we have this positive correlation between residuals, which suggests that if we just use Euclidean distance, we might overestimate some of the effect sizes observed. So, we use the Mahalanobis distance to correct for that.

This is a slide to try to explain the Mahalanobis distance and just forewarning, I'm not actually a statistician by training. But I think this is a good way to understand it, so let me walk you through it a little bit.

On the left we have a graph where the Y-axis is some hormone, B. The X-axis is Hormone A.

Then we have three screen concentrations of some imaginary chemical. These three screen concentrations are three points in this two-dimensional space defined by these two hormones, A and B.

This dotted line is the error distribution around these two hormone measurements. What you can see is that it is an ellipse. It's elongated in the Y direction, which means there is a greater error on the measurement of Hormone B than Hormone A.

If you look at the effect sizes of concentration one to three, and concentration one to two, you can see that they almost look equal if you just look at those vectors, but that actually concentration three is more standard deviations away from concentration one than concentration two, because it's breaking free of this error ellipse.

What we'd like to do is put this on a scale where the hormone concentrations are uncorrelated. They have the same standard deviations so you can see this effect size difference more easily. That's the graph on the right.

We have the rotated and scaled axis two, the rotated and scaled axis one. And you can see

now, concentration one and two and three as the points in that two-dimensional space. Again, the error distribution around the measure of these hormones represented by the dotted line -- which is now a circle, meaning that the standard deviation is the same for both hormones in this two-dimensional space. And now you can really see the effect size of concentration three is actually many more standard deviations away from concentration one than concentration two. Conceptually, this is what's done prior to computation of the Euclidean distance in this new space.

Unfortunately, we have an 11-dimensional space which is very hard to visualize. We've shown two dimensions, but really what we've done is Mahalanobis distance computation in 11-dimensional space where each axis corresponds to the natural log of the measured concentrations of one of the hormones in this analysis.

In brief, basically the degree to which variation among the replicates is correlated across hormones was estimated. The covariance matrix that characterizes both the noise variance and the correlation among hormone levels across replicates,

after taking into account the chemical concentration, was constructed. Computation of the mean Mahalanobis distance at each concentration of chemical was screened; and I'll show you the equation for that on the next slide.

This is the derivation of the mean Mahalanobis distance. Most of this equation just follows the form of the Mahalanobis distance derived by Mahalanobis in 1936. The one key difference is that we're calculating a mean Mahalanobis distance that's related to the number of hormones available at that chemical concentration.

Here in equation one you see the mean Mahalanobis distance, or mMd is equal to the square root of the transpose matrix of the natural log transformed steroid hormone concentrations at concentration c -- it's basically the transpose matrix of the fold change data using the DMSO as the control -- times the inverse of the covariance matrix estimate, times the matrix of the fold change values at that chemical concentration c. All divided by the number of hormones available.

From that, at each concentration of chemical, you'll have a mean Mahalanobis distance.

You'll have a set of mMd values from which you can select the maximum Mahalanobis distance, which is what we've done here. As the mean Mahalanobis distance generally increases with increasing concentration, a greater maxmMd should increase with increase in concentration of chemical and increased potency, i.e. activity at lower concentrations. This is true if the plot of the mean Mahalanobis distance by concentration is monotonic.

Just a little bit of detail on the covariance matrix estimation, really just a brief overview of how that was done. The estimated covariance matrix essentially characterized the noise variance and the correlation among the measured steroid hormone concentrations across replicates. The covariance matrix was computed for the multi-variant response, which ranged from nine to 11 steroid hormones per chemical screen.

If any data were missing, the hormone measure was dropped from that block prior to linear model fitting. This only affected one of the eight screening blocks performed, which included 81 chemicals, preceded with nine of the 11 hormones.

Then the unweighted average of those eight block specific covariance matrices was used as the full pooled 11 by 11 covariance matrix. So really, it's an estimate that was used in the mean Mahalanobis distance calculation.

I want to get to some more practical examples to show you how this worked. Here is an example of a chemical with moderate effects using this pathway approach. It's atrazine. On the left-hand side is a radar chart where each spoke of the radar chart is a hormone in our analysis.

In the middle there are some red dotted lines that indicate 1.5 or 1.5-fold, the vehicle control, just to give you some context. As the color of this line intensifies to become more blue, it's a higher concentration of atrazine.

What you can see is an effect largely on the estrogens and some of the corticosteroids kind of pulling the effect away from the center of the radar chart, showing you that it's really driving that response in a positive direction, exceeding that 1.5-fold vehicle control.

On the right-hand side is the compressed data. This is a plot of the mean

Mahalanobis distance by concentration of chemical. You can see here in the red dotted line, we've annotated the critical limit, which was simply derived so that you could have a threshold to discern an mMd that was greater than what would result from sampling noise. It's not necessarily biological criteria, not at all. It's really just bounding the sample noise.

And you can see here with increase in concentration of atrazine, you have increasing mean Mahalanobis distances and we've selected the maximum here, annotated with this blue box. This yielded a moderate adjusted maxmMd of -- that number is not correct. It should 3.14. I don't know why it says that.

Here is an example of a negative chemical, benfluralin. You can see here on the left, again, the same radar chart style where the spokes represent steroid hormones and intensifying blue concentration represents increase in concentration of benfluralin. You can see that all of the concentrations for benfluralin fall within that plus or minus 1.5-fold control. This provides an example of a chemical with a negative pathway result.

You can see to the right, the critical limit now is at the top of the graph because the Y-axis is different here. You have a distance here of about 1.8 being the critical limit. And all of the mean Mahalanobis distances, including the max, fall below that. So, this was negative.

Here is a third example, a strong effect of a chemical is mifepristone, also known as RU486. This chemical had very strong effects on progestogen, steroid hormone synthesis visualized here to the right on the radar chart. You can see, on the right-hand side of the slide, the mean Mahalanobis distance plot. Even at the lowest concentration tested, we have a mean Mahalanobis distance of about 10 and then the high adjusted maxmMd was not 171. It was, I want to say, 33 or something like that. Sorry, I don't know why the numbers are changed from my slides.

Those are just a few examples of kind of how this data flow worked from taking 11 steroid hormones to a compressed vision of using mean Mahalanobis distances, and then selecting that maximum as a number that could be used for prioritization.

Now I want to show you a couple slides that focus on different ways of looking at the data, some reproducibility, things like that. MaxmMd was generally reproducible and quantitatively distinguished chemicals with larger effects.

I want to talk a little bit about this graph with you. On the Y-axis is the maxmMd value and each dot is a chemical. Across the X-axis is a steroid hormone hit count. I thought this was an interesting way to look at the data. One of the initial questions that I received was, Katie, you got this Mahala-something distance, and could you just add the steroid hormone hit count up and that would give you relative priority?

Well, there are a couple problems with that. One is that you only have 11 numbers and you don't have enough range to distinguish chemicals. The other issue was it doesn't really quantitatively distinguish chemicals very well. I'll show you how the maxmMd does maybe a better job than that very simplistic type of view.

The first example is looking at these negative maxmMd values. You can see here on the lower end of the Y-axis there are these clear circles. The

clear circles indicate chemicals for which the pathway response was negative. And you can see in terms of steroid hormone hit count, they actually span zero to six. Mostly clustered zero to three, but one kind of hanging out over here.

And what this suggests is that you can have extremely borderline responses as annotated by an ANOVA that maybe on a quantitative basis are not that interesting. The maxmMd gives us a better way of distinguishing those really, kind of weak, borderline, maybe inactive chemicals from something that is stronger.

Bisphenol A is one of the reference chemicals in the OECD reference chemical set. But it's also a chemical often used as a reference in ToxCast because it's internally replicated in our library. It was actually screened three separate times, three separate chemical samples were used.

You can see here that actually for Bisphenol A, the maxmMd was pretty constant at about five. I've boxed in these values here that all kind of run across five. But the steroid hormone hit count ranged from five to seven. You see a very stable maxmMd, but kind of a variable steroid hormone hit

count. Again, that's part of the issue with binarizing data as positive or negative. It's to be expected.

Another example, two more OECD reference chemicals, looking at where they fall. So, EDS is ethylene dimethanesulfonate, and finasteride. They actually have the same hit count, but very different maxmMds. EDS is this inverted clear red triangle at the bottom. At the top, the orange circle is finasteride. Finasteride had a maxmMd exceeding 10, whereas EDS was actually a pathway negative. So you can see, we have the ability to distinguish a really strong effect from maybe a set of very weak effects, and quantitatively distinguish those chemicals.

More evidence looking at the reproducibility of the maxmMd -- and I think maybe Dr. Pullen Fedinick asked about this, or maybe Dr. Clewell. We actually did look at reproducibility using 107 chemicals of our set that were replicated in more than one experimental block. This subset of chemicals had maxmMds that ranged from one to 35, which covered most of our range for maxmMd. So, it's a pretty representative subset.

It's represented here by the graph. On the left-hand side are very, very tiny chemical names that you probably cannot read. On the X-axis is the maxmMd. You can see for each chemical, each sort of parallel line here, there is usually two or maybe three circles indicating that it was tested in two or three blocks.

The black solid circles indicated that there was a pathway positive, so a maxmMd that exceeded the critical limit for that chemical. And the clear circles, which are really clustered down in the lower left-hand corner, were pathway negatives using this approach, meaning that the maximum mean Mahalanobis distance was below the critical limit for that chemical.

What you see is that actually 88 percent of the maxmMd pathway response is replicated. That's just considering the pathway response is -- as binary, did it exceed the critical limit? So, 88 percent of the time, we could get the same response if we tested in that replicate block. And to contrast that, we also looked at the recall with our ANOVA logic, and the recall on that was only 65 percent. We

think that perhaps there is some value in this approach.

You can see here that I've annotated the residual standard deviation for this chemical set as 0.33. Actually, we've used that to derive a 95 percent confidence interval on predicting the maxmMd. Basically, it's plus or minus 1.5 units on the arithmetic scale, is what that comes out to.

I want to highlight one row that bothered me in particular, and maybe it stands out to you. There is this chemical here, 1,2,4-Butanetriol, where actually it looks like the standard deviation between those two replicates was pretty different. What we found was a sensitivity in our analysis.

So, this highlighted outlier demonstrated that the maxmMd is susceptible to missing data. The larger pathway response -- so the one shifted to the right, suggesting a higher maxmMd and a higher effect size -- resulted when much of the data for 1,2,4-Butanetriol was below the lower limit of quantitation. And we used a substitute basal value of LLQ divided by the square root of two.

Basically, that tells us that when we have missing data, we may inflate the effect size.

But that's going in a direction that maybe we can live with because it's a little bit of a false positive direction.

Finally, I want to bring it back to the OECD inter-laboratory reference chemical activity and show you some of the concordance for those chemicals and then how it lined up to the pathway response, and then how it lines up with the maxmMd quantitatively.

To the right here, we have a little geometric tiling figure, and all of the OECD reference chemicals where we had an overlap with high-throughput screening are listed. On the bottom you can see annotated HT and OECD -- again meaning high-throughput and OECD, of course. And then E2 for estradiol or T for testosterone, and then up or down. So these are the different effect types. And in terms of color, to try to make it easier, I've coded all of the estrogen effect types as yellow, and the testosterone as green.

Where you see two yellow blocks together, for instance for estradiol down, that means that we saw agreement across the high-throughput and the OECD inter-lab validation, as already annotated in those confusion matrices. The gray blocks indicated an equivocal finding. So, you'll find some gray

blocks in our interpretation of the OECD work where the labs didn't agree. Then clear just means no effect.

Again, this is the concordance we've already walked through, but what's new here is the pathway response, which is annotated with these blue blocks. Blue means that there was a positive pathway response, i.e. the maxmMd was greater than the critical limit. And the clear blocks mean that either it was negative or not tested in concentration response.

Here we've added the maxmMd. Here it's the log 10 of the maxmMd, and the chemicals are actually ordered by this value. So, the highest maxmMd was for mifepristone and the lowest for nonoxynol-9. I think what you can see here -- there are a couple of things.

One is that if there were findings in the OECD set, there was typically a pathway positive. Even in cases where we may have missed the particular effect type. So, mifepristone is a great example. You see a yellow block here, where in the OECD inter-laboratory validation, estradiol was decreased. We didn't have that finding in high-throughput

screening, but we did have a very strong pathway response, so we were able to pick that up.

The other thing that I think stands out here is that the stronger reference chemicals tend to have high maxmMds. So, mifepristone is an abortifacient drug. It's very potent in this biological space, and it comes out with a very high maxmMd. But genistein, for example, in the OECD inter-laboratory validation, increased estrogen responses by greater than 20-fold in some cases. It ends up being one of the stronger actors in our set as well. Ketoconazole, danazol, prochloraz, fenarimol, letrozole, and Asteride, these are all either pharmacologic agents or conazoles, and so they're well known to act on this pathway.

And then kind of this middle chunk of chemicals with moderate maxmMds are chemicals that maybe aren't necessarily pharmacological agents targeted to this pathway, but still have a reasonable response. Then negative chemicals like EDS were negative for the pathway response. I think in general, this really shows you how to pair together that concordance confusion matrix set and the pathway approach.

The biggest caveat to trying to do this, of course, is that the reference chemical effects on progestogen and corticosteroid or glucocorticoid hormones is mostly unknown. So, the pathway response will reflect any effect on the four steroid hormone classes; whereas, we typically don't have that information for all of the reference chemicals.

I'm happy to take questions on that meaty section before I move on.

DR. KRISTI PULLEN FEDINICK: I have a question. I just wanted to try to understand the recall, so for the ANOVA -- the comparison to the pathway and the ANOVA. So, you had 85 percent that were replicated across blocks, but then 65 percent -- does that mean that whether or not it was positive or negative in the pathway, you had 85 percent of concordance essentially?

DR. KATIE PAUL FRIEDMAN: Yes, just concordance. Not saying which direction. Just that they agreed.

DR. KRISTI PULLEN FEDINICK: And then 65 percent was saying that, so the 11 hormones could

change between those? So, it might not be the same 11 that you're seeing? Is that correct or no?

DR. KATIE PAUL FRIEDMAN: Let me think for a second. I lost you in your question. Let me rephrase a little bit. The 88 percent of the maxmMd pathway response is replicating, just means it was negative both times or it was positive both times. Whereas, the ANOVA recall, we lumped all of the steroid hormones together and just said, okay, was it either negative both times or positive both times. And then just looking at all of those responses together.

So, only 65 percent of those responses were concordant from block to block. That really highlights how you call a hit is really important.

DR. KRISTI PULLEN FEDINICK: Do the hormones replicate themselves? If you're looking at whether or not the same 11 -- do they have the same patterns across the blocks? Does that make sense? If you have --

DR. KATIE PAUL FRIEDMAN: Oh, yeah. Okay. That's a different question than I was thinking. In this case that's not what we were examining, but we could look at that to see which

hormones -- I mean, then again, you're back to the ANOVA logic to determine which hormones went up individually.

But you could look at, for instance, did the mean Mahalanobis distance that was at the critical limit -- so the threshold mean Mahalanobis distance -- what hormones were altered at that concentration, and were those same hormones altered in that concentration in replicate testing. I think that might be a way to answer that kind of question, and probably relevant if you were considering implementing it programmatically.

DR. JAMES MCMANAMAN: Other questions. Dr. Clewell?

DR. REBECCA CLEWELL: I know, I feel like I just keep asking all the questions. I want someone else to speak up. I'm going to admit that the Mahalanobis is way outside of my comfort zone, so I'm hoping we have some awesome statisticians on our panel that will take care of that.

When you look at the Venn diagram -- I'm going to contend that you sort of made this happen by forcing a cutoff of three. You don't have more than three hormones in any one of those hormone blocks

because two of them are dropped, right? So, when I went back and took those two out of the diagram of the hormones, three is your maximum number of hormones for any one group of -- you know, corticosteroids versus estrogens versus --

DR. KATIE PAUL FRIEDMAN: Right, there are some chemicals that we tested in multi-concentration that didn't affect any steroid hormone, for example. So, in this Venn diagram, only 628 out of 653 could even be charted because they didn't affect anything. We do have a set within the set that affected only zero or one or two; but most of the set affected three or more.

DR. REBECCA CLEWELL: Right. Most of the set had a cutoff of three. This Venn diagram is basically -- it kind of was forced to say that it affects more -- the chemicals generally affect more than one pathway.

DR. KATIE PAUL FRIEDMAN: Well, yes, yes. That's true. Except that if, for instance, a lot of these chemicals only inhibited aromatase and that was the predominant mechanism, then you would anticipate only an effect on four steroid hormones

within the estrogen and androgen class, and that's not what we really observed.

For the set that we tested, for this over 600 chemicals, it looks like things did more than just inhibit CYP19A1. Because that would have resulted in three or four hormones being modified, right, two androgens, two estrogens; but it looks like that wasn't the typical pattern.

DR. REBECCA CLEWELL: I would contend that maybe it would be a stronger case if we looked at the larger group of chemicals.

DR. KATIE PAUL FRIEDMAN: I agree.

DR. REBECCA CLEWELL: But my question around this is, then what's kind of the purpose of this? It seems that this idea that the hormones will generally be broadly affected rather than specifically effected, is sort of an important key point to how we end up doing the analysis, which is through this combined metric. Is that kind of the driving force here to say a combined metric is more appropriate because we expect these not to be very specific responses? Is that sort of the result of that logic?

DR. KATIE PAUL FRIEDMAN: I don't know, Woody, if you have thoughts. But my opinion on that

is that it's not really making a comment so much on the biology, I guess, of choosing a maxmMd as the metric. It's more if we don't understand the kinetics of how all of these enzymes are acting together in this implementation of the cell model, then the easiest way to compress those data is statistically and looking at the effect size across the pathway.

There may be valuable biological information in this graph -- because I think what I found surprising was that basically we didn't have a concentration of chemicals that only affected two classes. We didn't see most chemicals only affecting estrogen and androgen. We saw chemicals affecting much more than that. I'm not sure that biology is the only driver of picking the maxmMd as a potential metric. I mean, I suppose in an ideal world you would be able to develop a kinetic model. That requires a lot more data and time.

There is probably more than one way to solve the problem, but I think the biology that comes out in this Venn diagram that is really interesting is, if you think back to EDSTAC recommendations, I think there was a large emphasis on aromatase inhibition as being a very primary target. This

suggests that, at least in this cell model, that may be one of the targets, but that there are others as well.

DR. REBECCA CLEWELL: With the combined score, Mahalanobis -- I don't even know if I'm saying that right. I kind of have the same question that I did for the AUC, the combined AUC. What's the sort of driver of the final score then? Does that kind of weight multiple hormones over potency for a particular hormone? What's the kind of key driving characteristics that would push the score up?

DR. KATIE PAUL FRIEDMAN: There are a couple of scenarios, right? There are several scenarios that you could imagine. One scenario is you could have moderate or low effects on many hormones, and actually be able to see that effect across the pathway. You could have effects that maybe weren't significant on their own, but when you combine them you see a statistically significant effect size across the pathway.

The other scenario is you could have one or two hormones driving a very potent response. Mifepristone is a prime example of that where we only see a significant response for two hormones, but it's

so strong that it's one of the largest effect sizes -- largest maxmMds in the entire set.

I think there are multiple scenarios to drive maxmMd. It captures different scenarios from multiple hormones with borderline responses, to then get an effect that's actually meaningful, or just a subset of hormones that are driving it or maybe only one or two.

DR. REBECCA CLEWELL: One last time and then I promise I'm done. I don't want to harp on it, but I don't understand how mifepristone could have a high score if it only had two hormones, because then it would have not been tested in a dose response way, because there was a cutoff of three.

DR. KATIE PAUL FRIEDMAN: That wasn't the only criteria that we used for selection of chemicals. We also selected reference chemicals. We selected chemical classes of interest. We selected some chemicals that didn't meet that three or four steroid hormone criteria. There are multiple reasons that a chemical would be tested in concentration response and mifepristone was one of them.

DR. REBECCA CLEWELL: I would take that as an argument against that cutoff then, right? Okay. Thank you.

DR. KATIE PAUL FRIEDMAN: I guess just to respond to your final comment, I would maybe separate the idea of how you implement screening in an environment with limited resource versus program directed screening. So, those are two different questions.

The implementation used here I think was one way to do it, to maximize resource efficiency, but clearly not the only way. And if you had a list of chemicals that you knew a priori were of interest for various reasons, whether it was exposure or some other bioactivity, of course you would implement it differently.

DR. KRISTI PULLEN FEDINICK: That's a great segue, actually, to the question I had. I wanted to be able to try to see how to contextualize these numbers.

You mentioned phthalates and I looked at the phthalates, which I thought was actually an interesting class of chemicals to include in here. The maxmMd ranged from about -.25 to 1.1, for

chemicals that were analyzed in a recent National Academy of Sciences study on the application of systematic review for low dose hormone effects.

For DBP, so dibutyl phthalate, the maxmMd was 0.033. This is one that the NAS panel, the consensus panel, found had evidence of impacting steroidogenesis, or at least an endpoint that could be related to steroidogenesis. How do we contextualize a 0.033? That seems like it would be on the low end and we might cut that out.

DR. KATIE PAUL FRIEDMAN: I think depending on what number you're looking at, if it's the adjusted maxmMd, if it's greater than zero it just means that it exceeds the sample noise around baseline. It would be a pathway positive, but barely. If it's not adjusted, then it would possibly be negative. Adjusted simply means that we've subtracted the critical limit from the value. That must be adjusted that you're looking at.

DR. KRISTI PULLEN FEDINICK: It was the adjusted. Then that, say, DBP, would be marginally -- the phthalates essentially don't really work in this model?

DR. KATIE PAUL FRIEDMAN: Pretty weak.

DR. REBECCA CLEWELL: I'm sorry, I have to comment on that because I've done so much work with the phthalates over the years. The thing about the phthalates is, first of all, the active form is the monoester. So you wouldn't expect the diester to be positive in these assays.

The other point, which the NAS has been batting around and has been a big topic, is that there is actually a certain amount of evidence that the phthalates are not human relevant.

They are definitely steroidogenesis inhibitors in the rodent, but there is actually like xenograft models, so I wouldn't -- I'm not saying that there couldn't possibly be some effects here, but I don't think it's strange that they didn't see it with DBP.

DR. JAMES MCMANAMAN: These are supposed to be questions for clarification.

DR. REBECCA CLEWELL: I'm so sorry. I'll drop it.

DR. KATIE PAUL FRIEDMAN: That was very helpful. Thank you.

DR. GRANT WELLER: One quick clarification question. So, when you're estimating

this covariance matrix you describe covariance of the residuals. That's basically adjustment for different chemicals in different concentrations. Is that right? You're subtracting out predicted for those?

DR. KATIE PAUL FRIEDMAN: It's performed at each concentration.

DR. GRANT WELLER: My other question is on the Mahalanobis distance. There are certainly a number of different multivariate distance metrics, but Mahalanobis is really convenient in a number of ways when your sort of underlying data are normally distributed. Can you comment at all on any investigation of that, just empirically?

DR. KATIE PAUL FRIEDMAN: Do you want to talk about that, Woody?

DR. WOODROW SETZER: Not really. Yes, you're right. The form of the Mahalanobis distance, it sort of comes from multivariate normal distribution. Obviously, once you've got the form, you can use for it any distribution. You asked if we've explored the sort of multivariate normality of residual error. Not super extensively. I can tell you that it's sort of a crude look at residual plots and so forth, they look sort of vaguely univariately

normal anyway, one at a time. We've done no looks to check for multivariate normality at all.

DR. JAMES MCMANAMAN: Thank you. I think we can move on.

DR. KATIE PAUL FRIEDMAN: Great. I'll skip ahead. I'm coming up to the end, so hang in there with me.

Just to summarize what we've been through in the two parts of this talk. The high-throughput H295R screening assay was evaluated versus the OECD inter-laboratory validation low-throughput assay. We analyzed all of the high-throughput data using very similar ANOVA analysis and logic to what is in the OECD test guideline to enable that comparison and have it be more direct. That's summarized in the confusion matrices.

We've also performed a novel integration of the 11 steroid hormone analytes for a pathway level analysis using these high-throughput screening data. To summarize that, we've computed a mean Mahalanobis distance for each chemical concentration and screen. And then from that set of mean Mahalanobis distances for each chemical, we've

selected a maximum as a potential useful prioritization metric.

Regarding evaluation of the high-throughput screening assay and its performance, we had fairly good concordance of results with accuracies that range from 0.75 to 0.91 for effects on estradiol and testosterone. This is in contrast to agreement among the labs themselves in the inter-laboratory validation, which generally approached 90 percent. Minor disagreement between the high-throughput and low-throughput results occurred for chemicals typically that had perhaps borderline activity, or activity at really high concentrations.

A good example of this is 2,4-dinitrophenol which, in the OECD laboratory validation had some effects, but the actual potency of that chemical in their study ranged five orders of magnitude. Very large, from 0.001 to 100 micromolar for the lowest effect concentration. And so for that chemical, we only screened it at an MTC of 10 micromolar and found nothing.

You can see whenever you're doing screening, it is bound to be different. But these borderline activity or activity at really high

concentrations or really variable range, might be places where we have disagreement more often.

We would contend that the maxmMd might be useful for prioritization and weight of evidence applications. Calculation of the set of the mean Mahalanobis distance values reduced that 11-dimensional question to a single dimension. And the maxmMd appeared to provide a reproducible and quantitative approximation of the magnitude of effect on steroidogenesis within the H295R cell model. It quantitatively distinguished weak, moderate, and negative chemicals in this particular model.

Given mean Mahalanobis distance at each concentration, you could actually model, as I mentioned previously, the mean Mahalanobis distance at the critical limit. So, that threshold or lowest effect concentration could be modeled as a mean Mahalanobis distance.

That value, you could envision using as a concentration to review effects on specific hormones, as we discussed. So, if you were really interested in what hormones were affected by a chemical, you might select that more lowest effect concentration based on mean Mahalanobis distance and

look across and see what happened. It might be useful in that way, in a weight of evidence application.

Of course, limitations. And there are a lot, but specific to this I would say there are fewer reference chemical information available here than for other parts of the Endocrine Disruptor Screening Program. As soon as you move away from estrogen, it's a very steep hill down to having full reference chemical set information. In particular, in this case, information on corticosteroids and progestogens is typically lacking. Although there is information in the literature, and that could be scraped.

The potentially limited metabolic capacity of this assay. So, fortunately, H295R actually do express some xenobiotic metabolizing enzymes; but of course, they may not generate all of the relevant chemical metabolites, and that would have to be studied in greater depth. But, of course, as Richard Judson mentioned, there are efforts within EPA to retrofit some of our high-throughput screening assays, and also to predict metabolites that might be of concern. There is ongoing work to try to address that limitation.

The other limitation already mentioned by Dr. Judson in discussion with all of you, is the restriction to DMSO soluble chemicals. And of course, the Center is working to expand that to water soluble chemicals.

Finally, I want to acknowledge that with all the work that we do in the Center, there is usually a team and this was the team that worked on this. Derik Haggard, who is a postdoc in our group; Woody Setzer, who has joined me at the table; Richard Judson, and Matt Martin, and Agnes Karmaus, who have moved onto other endeavors, but contributed to this work. So, thank you.

DR. JAMES MCMANAMAN: Thank you. Any questions for this last part of her presentation? All right. Well then, I think we should take a break. Maybe 15 minutes. Be back here at what, ten after?

[BREAK]

DR. JAMES MCMANAMAN: Break time has come to a close. If there are issues related to the discussion that need to be read into the minutes during the break, please remember to do so.

Okay, Dr. Bever.

DR. RONNIE JOE BEVER: Hello, again. I will be presenting the high-throughput H295R assay discussion. And again, this is going to be structured similarly to my presentation after the AR discussion. I thank Dr. Paul Friedman because she has thoroughly described the high-throughput H295R assay and the use of the maximum mean Mahalanobis distance.

Once again, what I'll be doing --

DR. JAMES MCMANAMAN: Dr. Bever, could you put the microphone a little bit closer? Yes. They're just not picking up the way they should.

DR. RONNIE JOE BEVER: Okay. What I'll be doing is basically presenting how we're going to use this assay and why we believe it's ready for that use.

Let me specify right here the type of the high-throughput H295R assay we will be using. As a program, the Endocrine Disruptor Screening Program, we will ask for concentration response data. Dr. Paul Friedman has already shown you how this high-throughput H295R assay is, for the most part, an upscaling of the low-throughput H295R assay. As such, we feel like -- we'll be using it as an alternative.

We'll also be using it to prioritize chemicals for the EDSP testing.

Now, once again, I'll be reading some of these charge questions, but that's not for the purpose of really deliberating a charge question today. It's basically to present you with our point of view.

The first charge question says, can we use this high-throughput H295R assay as an alternative for the low-throughput H295R assay when we are measuring only testosterone and estradiol levels?

Now, I'm going to discuss the validation principles which kind of wraps up both -- two of the charge questions. First of all, the relevance. And once again, this maintains the mechanistic and biological relevance of the original. We're still measuring estradiol and testosterone, even when we're measuring nine other hormones in the pathway model.

The fit for purpose, I put up here 75 to 91 percent agreement; however, if you used the revised accuracy measures where the two problem chemicals are removed, you're getting 80 to 95 percent accuracy. It's important to remember that the

concordance in the OECD testing among the labs with the low-throughput assay, was 90 percent. We feel like this fit for purpose is a pretty grand score here.

Furthermore, using the pathway model, it helps to eliminate equivocal and discordant results. For instance, six potential false negatives were correctly identified with the pathway model.

Reliability. One-hundred-seven chemicals were screened in more than one screening block. Eighty-eight percent of these chemicals produced the same answer across block when using the maximum mean Mahalanobis distance. We say that this is a very good reliability. Dr. Paul Friedman also showed you the Z-primes, the Z-factor scores. They were typically excellent, showing that these assays are good and that the variability is acceptable.

Transparency. Once again, we tried to provide all the data necessary for people to understand the methods, as well as understanding the statistics that we performed to be able to reproduce those results.

Our second charge question with the steroidogenesis basically asks if the high-throughput

H295R assay measuring 11 hormones can serve as an alternative to the low-throughput H295R assay.

There are some advantages to measuring 11 hormones, and I think that Dr. Paul Friedman has really already highlighted these. But just to reiterate, when you're measuring 11 hormones it can be much more informative and robust than when you're measuring two hormones. The Venn diagram supports that actually more than estrogens and androgens are being affected.

It's useful for rapid prioritization using this maximum mean Mahalanobis distance. A pathway approach allows greater sensitivity. Once again, some of the potentially false negatives were correctly identified using the pathway model.

We are hopeful that this more informative and robust pathway model could be useful in identifying modes of action; but, as Dr. Paul Friedman said, this can be complex.

The last charge question, basically the last sentence here is asking for the comment on the strengths and limitations of the maximum mean Mahalanobis distance and the pattern of steroid hormone responses in the high-throughput H295R assay

for chemical prioritization and weight of evidence applications.

Here we feel that the maximum mean Mahalanobis distance is a very useful metric in this assay. It takes into account magnitude of effect. Basically, the maximum mean Mahalanobis distance provides a single numeric value to characterize the magnitude of effect on synthesis of 11 steroid hormones for a given chemical. It provides a quantitative answer. It's more than just a simple yes or no on an induction of a particular hormone. And this will allow prioritization.

In conclusion, we feel like the basic validation principles have been met. The Agency feels that measuring multiple hormones confers some advantages. And we contend that the maximum Mahalanobis distance is a useful statistical metric for this assay.

That's all I have. Thank you.

DR. JAMES MCMANAMAN: Questions for Dr. Bever? Yes?

DR. KRISTI PULLEN FEDINICK: You're using the word in here, "prioritization," especially for question four. Are you talking about using it for

prioritization before going into the Tier 1 tests?

DR. RONNIE JOE BEVER: Yes.

DR. KRISTI PULLEN FEDINICK: And then the separate question then in the prior two questions is whether or not it could replace a Tier 1 screen, right? For this question it's saying, can we use this, just overall, to say whether or not these chemicals should then go through any of the Tier 1 screens? And then go into Tier 2 if the evidence suggests such. Or are you saying -- can you clarify a little bit more about the word "prioritization" here?

DR. RONNIE JOE BEVER: What we're talking about, prioritization, is using all the high-throughput assays that we can -- basically what we're looking for is endocrine bioactivity information. If you remember the EDSTAC slide I showed at first, the very first block is, what do we know about the chemical. And if we know about its bioactivity, as well as exposure, we can make judicious choices on which chemicals to evaluate first.

The high-throughput assays have already been run on a few thousand chemicals. As for an example, the H295R assay, that's providing us with

endocrine bioactivity. We're going to use that in the prioritization stage.

DR. KRISTI PULLEN FEDINICK: I'm just going to clarify that again. I'm looking at the ESTAC conceptual framework now. This was your slide nine, I don't know if you should go all the way back to that, but -- so prioritization is something that would happen before the Tier 1 screen?

DR. RONNIE JOE BEVER: Yes.

DR. KRISTI PULLEN FEDINICK: For charge question 4, or the last one we looked at, it says, "using this high-throughput screen to prioritize chemicals." So, is what the Agency asking to say, can we use this high-throughput H295R assay to set the priorities before we would even send them through the AR assay, the ER assay, all of these other assays? You're going to use this one assay to prioritize and then move into the screen?

DR. RONNIE JOE BEVER: It's not one assay. We consider all the bioactivity information that we have. It's kind of like a screening risk assessment, so to speak. It's not a true risk assessment. It's not a weight of evidence thing in that we don't have complete information. It's using

the information we have to make good selection of the chemicals that we should test first.

We're not just going to use an H295R assay during prioritization. We're going to use H295R assay, ER pathway model, AR pathway model, or whatever other source of data that can provide us information on endocrine bioactivity. That's just one part. We want to look at exposure also.

DR. SEEMA SCHAPPELLE: Also to answer the second portion of your question, yes, this would be data that would be available as an alternative to the Tier 1 screen.

DR. KRISTI PULLEN FEDINICK: So, it's for both prioritization and for screening?

DR. RONNIE JOE BEVER: Yes.

DR. SEEMA SCHAPPELLE: Availability and alternative, yes.

DR. RONNIE JOE BEVER: That was the first slide of my mini presentation was, we're wanting to use it for two things. We want to use it for an alternative for the low-throughput H295R assay, and we want to use it in prioritization.

DR. JAMES MCMANAMAN: If there are no further questions, I think we can move on.

DR. RONNIE JOE BEVER: Thank you.

DR. JAMES MCMANAMAN: The next presenter is Dr. Scott Lynn.

DR. SCOTT LYNN: I am Scott Lynn. I'll be here to present to you the slideshow on Thyroid Conceptual Framework. I work in the Office of Science Coordination and Policy in the EACPD division. I've been sick, so I apologize now if I cough or if I come off as somewhat groggy as I'm presenting this.

This is an overview of the slideshow that I'll be presenting, and it mirrors very much what you have seen in the white paper. Section 4.1 of the white paper was really a very brief overview of the thyroid pathways. Section 4.2 presented the 15 MIEs identified by the EPA for thyroid based AOPs. Then put those AOPs into an AOP network for thyroid bioactivity. Section 4.3 had the current EDSP thyroid related endpoints, the Tier 1 and Tier 2. Also, the high-throughput assay status and then a prioritization ranking for those. And a thyroid framework coverage. Section 4.4 was a next steps and challenges. We have two charge questions that relate to Section 4.2 and 4.3 and those would be the middle sections of the presentation.

Thyroid hormones generally are considered to be iodothyronine T3 or tetra-iodothyronine or thyroxine, T4. The major circulating form is T4 and this is generally an 80 to 20 ratio in human adults. T4 is considered to be the prohormone, and T3 is the active hormone. In target tissues, the T4 is deiodinized to T3 where it is activated and binds to a receptor to initiate transcriptional pathways.

Thyroid hormone is conserved across all vertebrate species. Thyroid hormone in humans is the same as thyroid hormone in frogs or lower vertebrates or fish.

There are a number of roles of thyroid hormones. I'm going to touch upon a few. They regulate diverse processes. In adults, they control metabolic rate and thermogenesis. In fetus, newborn, and children, they can mediate many aspects of somatic growth and development. They are especially critical for nervous system development, neurogenesis, migrations, synaptogenesis, myelination of axons and the shifting of cells from a proliferation to a differentiation.

Thyroid hormone receptors are present

in a number of different tissues: pulmonary tissues, cardiac tissues, and obviously neuronal tissues where they play a crucial role. I'm going to briefly go through the thyroid pathway overview here, and then I'll get into more detail as I describe to you the 15 MIEs that we're presenting.

The thyroid axis generally is considered to begin in the hypothalamus with the synthesis in the secretion of thyrotropin releasing hormone. This travels to the pituitary where it stimulates the secretion and synthesis of thyroid stimulating hormone. These would be receptor based targets and what I'm going to do is, as I go through this, I'm going to show you -- we classified the 15 MIEs into four different processes. They were sort of binned into these processes just to make it easy to conceptualize it. I'm going to outline those four processes as I go through this.

Thyroid stimulating hormones stimulate synthesis and secretion of thyroid hormone, or T4, from the thyroid. There are a number of MIEs associated with this, and this is considered to be thyroid hormone biosynthesis.

Thyroid hormone is then released into

the bloodstream where it travels through circulation. There are thyroid hormone binding proteins in the serum and then there are also thyroid transporters on tissues. These have been binned into the thyroid hormone transporter category.

Then there is peripheral tissue metabolism. These are going to be associated with hepatic nuclear receptors in the liver or deiodination or sulfation or glucuronidation. So, those are all binned into the peripheral tissue metabolism category.

Finally, the effects of thyroid hormone are manifested through binding and activation of the thyroid hormone receptor and the target tissues. And then transcription of genes. Those, we've put into the receptor based targets also.

That's a brief overview of the thyroid pathway. I'm going to move on now to Section 4.2 where I describe the 15 MIEs and the AOP network for thyroid bioactivity.

The first process or bin of MIEs are the thyroid hormone biosynthesis in the thyroid. This begins with the sodium iodide symporter, which is a transmembrane glycoprotein which regulates iodide uptake into the thyroid follicular cells and is the

first step in thyroid synthesis.

Next would be the thyroperoxidase, which is an enzyme secreted into the thyroid colloid. And this oxidizes iodide ions for addition onto the thyroglobulin which catalyzes the formation of thyroid hormone.

Next is pendrin, and this is an anion exchange protein that mediates the efflux of iodide across the apical membrane of the thyrocyte.

Next is dual oxidase, which is an enzyme exposed to the colloid and this co-localizes with the thyroperoxidase and generates the peroxide necessary for thyroid hormone synthesis.

Lastly is the iodotyrosine deiodinase, which is a deiodinase enzyme in the apical plasma membrane of the colloid. This catalyzes the deiodination of the iodinated tyrosines which recycle the iodide within the thyroid.

The next processes or MIEs that are -- processes, are the thyroid hormone transporters. There are two categories here. There are the serum thyroid hormone binding proteins. There are three serum proteins: thyroxine-binding globulin, transthyretin, and albumin. These are responsible for

binding and transporting the thyroid hormones through the circulatory system to the target tissues.

The next set are the thyroid hormone membrane transporters. These are solute carrier gene family proteins. They transport thyroid hormone across plasma membranes. And these include the monocarboxylate transporters, MCT 8 and 10, and also the organic anion-transporting protein, or OATP1C1 in particular.

The next set of MIEs are under the thyroid hormone peripheral tissue metabolism bin or process. These include the iodothyronine deiodinase. These are the enzymes that deiodinate the T4 to T3 and activate it, or they inactivate the T4. There are three types. There is deiodinase 1, 2 and 3, and these function in a tissue specific and temporal manner. They modulate thyroid hormone homeostasis in terms of tissue responses.

There are also hepatic nuclear receptors and these mediate phase one, two, and three metabolism. And ultimately, the disposition of endogenous and exogenous chemicals. These contribute to thyroid hormone homeostasis. Two important ones are going to be CAR and PXR.

Next is sulfation and glucuronidation. These are important hepatic and nephritic pathways that regulate the thyroid hormone catabolism.

Finally, alanine side chain. There are alanine side chains of T4 and T3 that can be metabolized by oxidative decarboxylation or deamination.

Moving onto the last bin of MIEs. These were put into the process of receptor-based targets. TRH receptors are gene protein-coupled receptors. There is TRHR1, TRHR2, which are primarily in the pituitary. They control the synthesis and the release of the thyroid stimulating hormones, as I said.

Thyroid stimulating hormone receptors are also GPCR primarily on the thyroid epithelial cells, and they control the production of thyroid hormones.

Thyroid hormone receptors, the nuclear receptors. They are activated by T3, the active form of thyroid hormone. These initiate regulation of gene expression in a wide variety of cell types, and there are thyroid hormone alpha, with one and two subtypes; and beta, with one and two subtypes. These show a

tissue specific and a temporal function in terms of life stage.

Then there is thyroid hormone transcription. Many thyroid hormone signaling pathways are mediated by transcription of the thyroid hormone receptor responsive genes, and these are critical for normal development and organ system functioning.

Heterodimerization with retinoid X receptor is necessary for transcription, and there are also a number of cofactors and coregulators that influence thyroid hormone transcription.

Now I want to introduce adverse outcome pathways. The adverse outcome pathway is a concept that's being utilized within the US EPA as a framework for organizing knowledge. Not only within the US EPA. I also need to point out, as I will later, that this is an international effort in terms of utilizing adverse outcome pathways. It provides a framework for organizing the knowledge about the progression of toxicity events across scales or a biological organization or hierarchical levels across biological organization that lead to adverse outcomes, either at the organism level or at the population level. These

adverse outcomes are the endpoints that are relevant for risk assessment.

The adverse outcome pathway approach begins with information about toxicant. And these are physical chemical properties of the toxicant, in terms of chemical category or profiles of the chemical.

Then from the biological perspective, it begins with the molecular initiating event, or MIE. This is where the chemical initially interacts with a protein or DNA. A protein could be a receptor, it could be an enzyme, it could be a transporter. There could be any number of key events along an adverse outcome pathway. These can be cellular responses or organ responses.

Some examples here are gene activation, protein production; secondary messenger changes in cells is another possibility. Organ responses could be altered function of an organ or altered development of an organ. Ultimately, there are adverse outcomes at the organismal level. These could be mortality or lethality.

One of the things that our program would be particularly interested in would be reproduction, obviously. Then population level, you

can have changes in population in structure. Or extinction of species could also be an adverse outcome.

This is the approach, the framework that the EPA is utilizing for a number of efforts. This is the approach that we have taken in terms of presenting the MIEs and thyroid conceptual framework to this panel.

What I'm showing you here is an example of an adverse outcome pathway. This one is thyroperoxidase inhibition. This is published in the OECD AOP Wiki. It is, as of right now, the only thyroid related AOP in the AOP Wiki. It is AOP number 42.

It begins with the methimazole and also propylthiouracil, are the two chemicals listed for the toxicant. Obviously, there could be a number of chemicals that would initiate this adverse outcome pathway.

These are known to decrease the activity of thyroperoxidase, which ultimately ends in a reduction in circulating serum thyroid hormone levels. This causes alteration in brain development, and the adverse outcomes are going to be

neurodevelopmental deficits, such as hearing loss, cognitive defects, neuroanatomy defects, synaptic dysfunction, or neuroplasticity defects.

I'm going to take some time to walk you through this figure. This is essentially the figure from the white paper. But what this does is this takes the 15 MIEs that have been proposed by the EPA, and it applies them to an adverse outcome pathway for a thyroid network. It's sort of linking together all of these MIEs and interrelated key events, and mapping them out to a number of different adverse outcomes.

The black boxes are the MIEs. These are grouped into the gray boxes. And these gray boxes generally represent the different thyroid pathway processes that I spoke about before. Receptor ligand interactions here and also up here in terms of the hypothalamic pituitary feedback. The thyroid here has MIEs that are associated with thyroid hormone biosynthesis. Down here we have thyroid hormone transport and metabolism lumped together. And then thyroid hormone catabolism and excretion here.

The light blue boxes are the key events, the KEs. These are ones that have been identified along the thyroid AOPs, and are presented

here. These are taken from the literature, in many cases, to establish these as key events associated with these adverse outcome pathways.

Finally, we have the dark blue boxes over here that are the adverse outcomes. It is indicated whether the evidence is manifested in mammals, amphibians, or fish. In some cases these are very specific in terms of teleost swim bladder inflation. That is going to be a specific endpoint for fish. Metamorphosis is a process that can occur in fish or in amphibians, so that is a potential for both. Up here we have rat thyroid tumors and also, neurological and cognitive impairments. And also, auditory impairments.

There are a couple things I want to do with this figure. The first is I want to highlight the importance of serum T4 in terms of being a key event that falls along many of the adverse outcome pathways that you can see here. But I also want to point out that not all adverse outcome pathways, beginning with the molecular initiating events that we've identified, go through this key event of serum T4 changes -- or changes in concentration of serum T4.

I should also point out here that the

aqua boxes, the hypothalamus and the pituitary are actual organs. And these are representative of circulating levels or hormone levels of TSH and TRH.

This is going to be the summary slide for the Section 4.2. I've identified the 15 MIEs across the different processes in the thyroid network. We've linked these MIEs to key events in adverse outcomes for putative adverse outcome pathways. And then we've summarized these into a chemically induced thyroid disruption network. We think that this is critical for us in terms of developing a thyroid conceptual framework.

The charge to the panel -- which I'll get deliberation on; I'm not meaning to bring this up as a question right now -- is to comment on the completeness of what we've presented here in terms of the number of MIEs, the different MIEs, are the key events comprehensive, and the adverse outcomes and just some feedback on the thyroid AOP network that we've presented.

Now I'm going to move onto Section 4.3 where I discuss the screening and assay status. First, I'll walk through the Tier 1 and Tier 2 thyroid related endpoints within the Tier 1 and Tier 2

battery. Then I'll present the high-throughput assay status and our prioritization ranking. And then, finally, the thyroid framework coverage that is present within the EDSP right now.

This is a slide very similar to one that was shown, I believe, by Dr. Schappelle earlier at the very beginning. Or maybe it was Dr. Bever. I don't remember. Sorry. This is the EDSP assays and the status of alternatives. As was mentioned before, in December of 2014 there were alternatives that were accepted for the ER model in terms of estrogen receptor, ER binding, ER transactivation, and uterotrophic.

At this meeting now, there are alternatives that are being presented for the AR binding and also for a high-throughput steroidogenesis assay. But what I want to draw your attention to are the thyroid related endpoints, or the assays that have thyroid related endpoints.

I've highlighted these in red for you. Within the Tier 1 battery, we have three assays: the amphibian metamorphosis assay, the female rat pubertal assay, the male rat pubertal assay. And then, for the Tier 2 tests, which are longer term life cycle or

multi-generation tests that are useful or utilized within a risk assessment framework, we have the avian multi-generation reproduction test, the larval amphibian growth and development assay, or the LAGDA as it's termed, or the rat assays -- this would be a two gen for a rat, or an extended one generation reproduction assay -- reproduction test or the EOGRT.

What I want to do is I want to briefly walk through those and describe the endpoints that are thyroid related or thyroid specific endpoints, but also mention potential thyroid related endpoints within each of these assays.

This is a table taken from the white paper. These are the Tier 1 test guidelines. The amphibian metamorphosis and the male and female pubertal assays. It shows the species. For the amphibian it's xenopus laevis. And then rat for the pubertal assays.

Here are thyroid specific endpoints. For the amphibian assay, we have endpoints such as hind limb length, developmental stage, and asynchronous development. They are really endpoints that are at the organismal level and could be indicative of potential population effects, too.

Then there is thyroid histopathology, and there are a number of endpoints within the histology that is performed. And, of course, there are thyroid related endpoints that would be tied to growth, body weight, or snout and vent length.

For the pubertal assays, we have organ weight, thyroid weight, and also measurements of serum concentrations of T4 and TSH. Then thyroid histology endpoints. And this is the same for both of the pubertal assays. Again, thyroid related endpoints are going to be tied to growth, body weight.

For the Tier 2 test guidelines, we have the avian two-generation toxicity test which was performed in Japanese quail. The thyroid specific measures are going to be thyroid size or thyroid weight. And circulating T4 concentration, along with T4 concentration within the thyroid. And egg yolk T4. Then thyroid histology. There are a number of endpoints there. Again, thyroid related measures are going to be growth, body weight.

The LAGDA is also with xenopus. It is a longer assay than the AMA, but the endpoints that are thyroid specific are very similar. Essentially, for the most part the same. Developmental stage,

asynchronous development, the time to NF Stage 62 -- which is a metamorphosis -- and then also, thyroid histology endpoints.

Finally, we have the EOGRT, where again organ weight, circulating concentrations of T4 and TSH, and a full thyroid histology list of endpoints. For this one there are optional thyroid related measures that would be neurohistopathology, neurobehavioral tests could be also measured or performed, and brain weight is also an endpoint that could be thyroid related.

What I have here is sort of an overview. I want to walk you through this. Across the top we have different types of studies. Here we have in silico and in vitro studies. This would be chemical categories, quantitative structure, activity relationships, or in vitro data that would be associated with high-throughput assays.

Here we have the three Tier 1 assays I just mentioned. The two pubertals, the AMA. Here we have the Tier 2 tests, EOGRT, LAGDA, and quail. Along the left we have endpoints that can be or are measured in each of these different tests. These endpoints range on the AOP -- and this is on the right -- we're

indicating where these endpoints would fall on an adverse outcome pathway framework.

What we see is that we have biochemistry in terms of measuring T4 and TSH and circulatory. This is going to be a cellular response. Thyroid weight is going to be an organ response. Histopathology, also an organ response. Changes in terms of metamorphosis would be an organism adverse outcome. Then ultimately, development could be tied to population level effects.

What I want to point out is that I've highlighted this MIE line -- row -- here for the molecular interactions. And what we see is the present Tier 1 and Tier 2 battery does not have any assays or any endpoints that specifically interrogate the molecular initiating event of a chemical on the organism. This is where high-throughput in vitro assays are very useful, and I would almost say imperative for us to get that data on MIE interaction, allowing us to map a full complete adverse outcome pathway for certain chemicals.

Now what I want to do is I'm going to walk through what was presented in the white paper in terms of high-throughput assay status. There were

four categories. The first one was existing. For this category there were one or more high-throughput assays that existed in ToxCast or Tox21. For these we had the thyroperoxidase, which falls under biosynthesis. There are two cell-free enzyme inhibition assays within the ToxCast/Tox21 platform.

For the metabolism, the hepatic nuclear receptors. There were multiple assays available, but I wanted to make note of the constitutive androstane receptor, CAR, and the pregnane X receptor, PXR.

Other MIEs for which they were given an existing classification under the receptor-based targets. The thyrotropin releasing hormone receptor. There is a GPCR cell-free receptor binding assay. TSH receptor. There is an agonist and an antagonist cell-based receptor reporter assay. These assays, I'll be showing the list of them, but they're also in the white paper and information on them can be downloaded by following the reference for the US EPA 2015 reference link.

Thyroid hormone receptor. There is one cell-free receptor binding assay. And then for thyroid hormone transcription, there are four cell-based thyroid hormone receptor transcriptional

reporter assays that cover alpha and beta. But I also want to point out that there are six cell-based retinoid X receptor for RXR alpha, beta, and gamma. And those are transcriptional reporter assays.

Moving on to the category, term developing. For this category there were one or more high-throughput assays that are presently being developed within the ToxCast or Tox21 platform. But the assay information of results in terms of chemical screening have not been made public yet.

The first one under biosynthesis is the sodium iodide transporter. This assay has been published, Hallinger, et al. The reference is in the white paper. This is a cell-based radioactive iodide uptake assay. That one, there is a follow-up publication that has done the Phase 1 ToxCast chemicals. And that publication is in review. I'll show that reference in a second.

Under the metabolism, there are three cell-free enzyme inhibition assays that are focused on deiodinase 1, 2 and 3. I believe there are three publications associated with that that are in different stages of development. None of them are available yet.

Then we get to the two categories where there is -- the next category, I should say, is promising. For this one, we were able to find and identify high-throughput assays in the peer-reviewed literature. These assays have potential. They are either high-throughput or are amenable to high-throughput, but these haven't yet been incorporated or developed into a ToxCast or Tox21 assay. There are a number of steps associated with that, so none of these assays have gone through that.

There are two MIEs for which this is the case. They are both in the thyroid hormone transporter bin process. The first is the serum binding proteins and this is outlined in Marchesini 2006. That was a cell-free binding inhibition assay on a biosensor chip. Then there are membrane transporters. This is a very recent publication -- I believe it came out in July or August -- for cell based T3 uptake assay with iodine detection. This is Dong and Wade of 2017.

The last category are the categories for which we could find no existing high-throughput assays. For the MIE, or any assays that are present that would interrogate this, would need basic research

and development in order to even begin to bring them into a high-throughput platform. Most of these fall within the biosynthesis: pendrin, DUOX, or the iodotyrosine deiodinase. Also, the peripheral tissue metabolism: sulfation and glucuronidation and the alanine side chain.

Next I want to talk about the prioritization rankings for the MIEs that were applied in the white paper. As mentioned in the white paper, these ranks were assigned based on a combination of biological relevance of the MIE in terms of the thyroid pathways, what toxicological evidence was available within the peer reviewed literature, and then also what was considered was the status that I just mentioned -- the assay status or the availability of assays in terms of their availability to be implemented into a high-throughput platform for ToxCast or Tox21.

The prioritization rankings for MIEs consisted of three levels. These were high, medium, and low. For the high prioritization MIEs, there were two within the biosynthesis and TPO. For the NIS, this is a developing assay as I mentioned. The Wang, et al. is in review in terms of applying the NIS assay

that is outlined in the Hallinger publication.

The thyroperoxidase is published. This is Paul 2014, Paul Friedman 2017, and the assay names are listed there. You can get details, annotated details on these assays if you follow the link to the EPA 2015 reference.

Then the deiodinase assays, which are in development, which I mentioned. The hepatic nuclear receptors for which assays do exist. And then sulfation and glucuronidation under the peripheral tissue metabolism.

The medium prioritization MIEs include those under the thyroid hormone transporters, the serum thyroid binding proteins, and the membrane transporters, which both were promising for the status, the high-throughput status. And then also, the receptor-based targets for TRH receptor, TSH receptor, and then thyroid hormone transcription.

One of the things that I want to point out is that there was an error in the white paper. Unfortunately, the assays for thyroid stimulating hormone receptor, in the process of putting in the links, got copied for the thyroid hormone transcription. The two assays here in red are the

proper assays. Those are the assays you want to look for if you go to the 2015 reference and download the Excel spreadsheet.

Finally, the MIEs that are ranked with a low prioritization include two within the biosynthesis, DUOX and iodotyrosine deiodinase, the peripheral tissue metabolism MIE of alanine side chain; and then lastly is thyroid hormone receptor binding, which is under receptor-based targets.

Now, I want to spend some time here. This is the initial conceptual framework for screening for chemicals that would have thyroid bioactivity. When you look at this figure -- this is in the white paper. On the left, we have the black boxes. These again are the MIEs. Up above is just the general overview of what we're seeing. We have chemical disposition here with parent molecule or metabolites. This would fall under the toxicant, the blue box, within the adverse outcome pathway graphic.

Here we have the MIEs. All the black boxes represent MIEs, which I have just presented to you. The light blue boxes, again, represent the key events. These are key events that are specific to the EDSP Tier 1 and Tier 2 assays. These are the key

events that are endpoints within the assays that I just presented. And then the dark blue boxes are the adverse outcomes that, likewise, are measured within the Tier 1 and Tier 2 assays.

The MIEs that are listed are only those MIEs which the EPA identified as medium or high priority. So, the MIEs for which we did not identify, or which were identified as being a low priority, are presently not on this framework. The orange outlines represent MIEs that are either existing or in development within the ToxCast or Tox21 framework.

What I want to do is I want to highlight a few points here. Again, I want to draw your attention – so, first off, this framework and the adverse outcome pathway network that was presented under Section 4.2, both of these do show feedback loops within the thyroid network or the thyroid pathways. We're considering feedback in terms of thyroid hormone effects on the hypothalamus and also on the pituitary. And also, feedback in terms of serum concentrations of T4 influencing tissue concentrations of T4 and vice versa, especially in respect to deiodinase inhibition.

One of the points I want to make here

is that serum T4 concentrations cover a number of potential adverse outcome pathways, as I mentioned before. It is the main key event within the EDSP Tier 1 and Tier 2 assays. The key events though that we do have do not cover all potential adverse outcome pathways. For example, deiodinase inhibition -- which would work at the tissue level -- this is a potential key event here for which there is no endpoint within the Tier 1 or Tier 2 assays.

We have adverse outcomes in terms of impaired metamorphosis or thyroid histopathology endpoints, or perhaps impaired neurodevelopment in mammals; but, presently, we don't have key events that fall upon that particular adverse outcome pathway.

Again, I should point out these boxes represent the different compartments of the thyroid pathway here; serum plasma, the thyroid gland, and the feedback, and also liver and target tissues for the MIEs.

To summarize this section of the white paper, we covered the EDSP test guidelines for Tier 1 and Tier 2, presented the endpoints that are within those. We overlaid those endpoints on the thyroid AOP network. We summarized the ToxCast high-throughput

assay status and the prioritization ranking that the EPA has presented for assay development. We have an initial framework for screening based on the thyroid AOP network.

The charge to the panel associated with this section is to comment on the prioritization ranking applied by the EPA for high-throughput assay development, and on the completeness of the thyroid screening framework -- the conceptual framework that we've presented to you.

The last section here is for Section 4.4 of the white paper. This is next steps and challenges. The goal of this section was to indicate to the panel that the EPA is considering or understands that there are going to be next steps and challenges associated with those next steps, and without committing to any specific method of approaching them, has outlined a number of possibilities and just identified -- or tried to identify these challenges.

The first is development and refinement of additional assays. As I showed you with the thyroid conceptual framework, we do have potential putative AOPs for which we will have, potentially, a

molecular initiating event and an adverse outcome, without key events. There is the possibility of developing assays that could fill those gaps for key events. That's essentially what we're asking the panel to give input on this week. Also, the high-throughput assays could be developed and are being developed, hopefully, to provide better coverage of the thyroid AOP network.

Next is going to be identification of reference chemicals. This is true for not only new assays, but also for extant assays. As has been presented earlier, reference chemicals must span a dynamic range of potencies. They need to be specific and sensitive, and they need to cover all potential modes of action in order to give confidence in the assays and in the assay results.

Next is development of performance- based approaches. This is highlighted in Section 1.5, which was covered by Dr. Bever, and ultimately is based on OECD guidance document 34.

Then development of an integrated strategy for analysis of assay data. One of the points that I want to reiterate that was written in the white paper, is that a single assay hit in a

high-throughput assay should not necessarily lead to a Tier 1 test order. Hopefully, we will have a number of assays -- either high-throughput or otherwise -- that will provide evidence of a potential initiation of an adverse outcome pathway to support any test orders that would go forward.

Compartment or other physiologically based models can and are being developed by some people within EPA, some researchers. Ultimately, we understand that we need to develop a more comprehensive framework for prioritization and screening of thyroid active chemicals.

That's the end. I don't have an acknowledgement slide, but I want to acknowledge people who helped generate the thyroid white paper. And that's going to be Dr. Bever, Dr. Andrea Kirk, and Dr. Katie Paul Friedman.

Next, I guess we just go to questions. Thank you.

DR. JAMES MCMANAMAN: Thank you very much. Thank you, Dr. Perkins for leading this. So, any questions for Dr. Lynn? Yes, Dr. Androulakis?

DR. IOANNIS ANDROULAKIS: I just had one question. I'm a dynamics guy, so whenever I hear

somebody talk about feedback, I get pretty excited. If the HPT axis is anything like the HPA axis, which most likely it is, this feedback that you mentioned, that you indicated, becomes very important. As you think about developing assays, the only way that you can really characterize and understand this, is if you can actually generate dynamic data.

When you talk about assays and things like that in this context, are you also envisioning the likelihood of generating some kind of dynamic data that will be able to give you some information regarding the importance of this feedback? Or there would still be the exposure outcome kind of --

DR. SCOTT LYNN: Can you clarify what you mean by dynamic data? I think the answer is yes.

DR. IOANNIS ANDROULAKIS: Temporal.

DR. SCOTT LYNN: Yes.

DR. JAMES MCMANAMAN: Other questions? Dr. Ehrich.

DR. MARION EHRICH: A question on -- you're talking about use of therapeutic agents maybe as reference chemicals. Methimazole is an anti-thyroid drug commonly used, but you don't see all those adverse outcomes from that because you're

treating a disease.

DR. SCOTT LYNN: I don't see why pharmaceuticals couldn't be used as reference chemicals. Certainly, I would guess that there were some that have been used for the other models. If it fits the classification for what we would be looking for, for a reference chemical. Does that answer your question?

DR. MARION EHRICH: It was just that when you're using it therapeutically, you had this big long line and you're going to have these neurological defects by the end, and yet you don't see those when it's used therapeutically.

DR. SCOTT LYNN: Well, I think it would be a matter of potency and dose. Does that answer what you're --

DR. MARION EHRICH: Or the disease of the animal or person that's taking the drug.

DR. SCOTT LYNN: Yes. Right.

DR. MARION EHRICH: I was just wondering what your thought was on the use of therapeutic agents for a reference chemical. That's going to be a big problem I see here for your thyroid development assays.

DR. J. DAVID FURLOW: I guess to follow up on that, in an AOP though, it seems to be -- especially if it's neurodevelopmental -- AOP, right? That's the adverse outcome. When the exposure happens is built into that. That's correct, right? So, an adult taking it will have -- methimazole -- will get a goiter, right?

DR. SCOTT LYNN: Yes.

DR. J. DAVID FURLOW: But the motor will be fine, right? So, it's a matter -- is that always explicit in AOPs? That the critical windows, that these are all built in, they're elaborated whenever they're laid out. They incorporated the -- I've seen some of these things written up, but that's always in there. Especially when we're talking about neurodevelopment, this could be true for AOPs for estrogens too, right? So, critical periods are essential; so, when the exposure happens has to be part of that AOP.

DR. SCOTT LYNN: Absolutely, yes. In terms of -- what I would say is that AOPs are going to be specific for a certain life stage or a certain situation, right? That needs to be considered, yes.

DR. JAMES MCMANAMAN: Any more

questions? Okay.

DR. STANLEY BARONE: Just to add to Scott's comments. Again, whether we're talking about outcomes, the molecular initiating events in this framework can be common, as you saw in the graphic. But the adverse outcome in a developmental context for fish or frogs may be very different than for mammals.

So, again, if you look at the framework, you may be measuring very similar molecular initiating events, but the outcomes and the life stage are going to vary across species. There will be additional AOPs for each species for each outcome.

DR. KRISTI PULLEN FEDINICK: Maybe you mentioned this, but it's after lunch and getting late. Was there a reason for wanting to prioritize the MIEs? Is it a resource question? Is it a scientific question? What's the reasoning behind wanting to exclude the low priority ones from the AOP network that you would use for screening? And why would you prioritize them overall?

DR. SCOTT LYNN: To answer your question, yes. It is a resource, and because it is resource, you want to approach it from a scientific perspective. It would be great if we could develop

assays immediately for everything, but that isn't really how it happens. So, we have to focus on what we feel is going to be the most important, and that's what we were trying to get across with the prioritization ranking.

DR. KRISTI PULLEN FEDINICK: Could you anticipate then, incorporating the low priority assays or MIEs or developing those over time as you grow this program out? Or do you think that those low priority MIEs will never be included in this network?

DR. SCOTT LYNN: That's a great question, and I can't give a definitive answer on that yet. One of the things I think is that we have to examine for more toxicological evidence for certain MIEs and determine if they are MIEs that would be responsive to the chemical universe that we're interested in.

DR. JAMES MCMANAMAN: Other questions?

DR. J. DAVID FURLOW: Sorry to follow up on questions. I have my own questions, too, but these are triggering ones I had. I think it is an important point to think through this going forward. If you have the high and medium and low priority targets, what does that mean? It would be nice to

know what that means functionally.

This will be an iterative process, I assume. So the idea is that you build out the framework, you build out your assays, you screen and you see who fits the biology or the responses. And then if it doesn't, you look at other targets? What's the sort of process?

DR. SCOTT LYNN: Exactly. It will be an iterative process in terms of what you just said. My understanding is that we're going to build out the assays that we've presented here, begin to develop models that would help us to prioritize and screen -- as we've shown with the other pathways -- and then ultimately, revisit and see where we might be missing things.

DR. STANLEY BARONE: If I may add onto that, to Scott's point. Just to remind the panel, in the 21-year history of the EDSP program, this has been an iterative, learning-by-doing program. The whole pivot to high-throughput and computational approaches, as you witnessed today, we showed that we have taken into account public comments and the previous SAP and recommendations in improvements of the current screening approaches. So, that will be part of the

effort.

DR. JAMES MCMANAMAN: That was Stan Barone. Other questions? Do you have another one?

DR. J. DAVID FURLOW: Can I follow up?

DR. JAMES MCMANAMAN: Sure.

DR. J. DAVID FURLOW: As part of this iterative process then, the idea is that can you also learn from the estrogen and androgen results? If you need 11 androgen receptor assays to replace one binding assay, or if you need 18 estrogen receptors, do we need to do that for all of these things, do you think?

What's sort of the minimum number of assays with different kind of readouts that one might need to feel confident that you've hit that MIE or you've -- I've got one assay, one assay here, four assays here. What have we learned from the estrogen and androgen stories? Or is it going to be totally different for thyroid, do you think?

DR. SCOTT LYNN: I wouldn't say that it's going to be totally different. I think there are absolutely lessons learned. I think you bring up a very good point.

For TPO, there were two assays, and

they were developed in order to build confidence in the results, if you look at the publications. I would say that the feeling is, is that that is how the process is going to move forward for all of these MIEs, where possible trying to identify and build orthogonal assays that will give the confidence to be able to ascertain if any given chemical is actually acting through that molecular initiating event.

That's not necessarily an easy task, though, for all of these. That's part of why we're coming here now to the panel, to ask for input on -- your recommendations on the way to do that.

DR. JAMES MCMANAMAN: Other questions? If not, then I think that we can adjourn for the evening. I'd like for all the panelists to meet in the break room for a brief post-meeting meeting. Then we'll see you all back tomorrow morning at 9:00 a.m.

[MEETING ADJOURNED FOR THE DAY]

* * * * *

DAY 2

DR. TODD PETERSON: Good morning. We'd like to start our session. One thing I forgot in my prepared statement yesterday was to kindly request if you have a cell phone to put it on vibrate. That would be nice.

This is Day 2 of our meeting on Continuing Development of Alternative High-Throughput Screens To Determine Endocrine Disruption Focusing On Androgen Receptor, Steroidogenesis, and Thyroid Pathways. I don't really have anything else to make comment of. We do have public comments this morning. We may have a little extra time left over from the comment period, if it's not used in which case we may move our agenda forward a little bit. And we'll see how that goes.

I'll turn the meeting over to the chair, Dr. McManaman.

DR. JAMES MCMANAMAN: Good morning. We'll start with going around and doing introductions again. I'm Jim McManaman at the University of Colorado.

DR. DANA BARR: I'm Dana Barr. I'm a professor of Environmental Health at Emory University.

DR. MARION EHRICH: Marion Ehrich, Virginia Tech Pharmacology and Toxicology.

DR. SONYA SOBRIAN: Good morning. Sonya Sobrian, Howard University College of Medicine Department of Pharmacology.

DR. SUSAN NAGEL: Susan Nagel, University of Missouri, OBGYN and Women's Health.

DR. TOM ZOELLER: Tom Zoeller, University of Massachusetts at Amherst, thyroid, hormone action and brain development and chemicals that interfere with that.

DR. GRANT WELLER: I'm Grant Weller. I'm a senior scientist at Savvysherpa in Minneapolis, Minnesota, and I'm a statistician.

DR. KRISTI PULLEN FEDINICK: Kristi Pullen Fedinick at the Natural Resources Defense Council.

DR. ED PERKINS: Ed Perkins, Army Corp of Engineers. I'm an eco-toxicology toxicogenomics, and hazard assessment.

DR. REBECCA CLEWELL: Rebecca Clewell, ScitoVation, in vitro and in silico tools to improve chemical safety assessment.

DR. MICHAEL PENNEL: Michael Pennell. I'm an associate professor of biostatistics at Ohio State University.

DR. IOANNIS ANDROULAKIS: Ioannis Androulakis, professor of biomedical engineering, Rutgers University.

DR. SCOTT BELCHER: Scott Belcher, professor of biological sciences and member of the Center for Health and Human Environment at North Carolina State University.

DR. VERONICA BERROCAL: Veronica Berrocal, associate professor of biostatistics at the University of Michigan.

DR. DAVID FURLOW: David Furlow, professor of neurobiology, physiology, and behavior at the University of California, Davis.

MR. TODD PETERSON: And my coffee's just sinking in. I'm Todd Peterson. I'm the DFO, Designated Federal Official for this meeting. And to my right is Tamue Gibson, who is my colleague also a DFO for the meeting.

Dr. David Jett is going to be calling in today, and so you don't see him physically present, but he's dialing in now. And we're expecting Dr. Shaw

to show up at some point. He's in travel coming to the meeting.

DR. JAMES MCMANAMAN: Okay. First, if we could, there's a special slide. Can we get that pulled up?

Scott Lynn was nice enough to send around a picture yesterday from the SETAC meeting. Many of you remember Steve Klaine, who was the previous chair of this committee. He passed away 2016. It's hard to see this I guess, but that's -- blow it up -- that's his pedigree. Those are the students that he was involved in training.

And those of you who don't know Steve and those of you who do know Steve realize that he was an exceptional individual. I mean both as a scientist and as a person. He had an exceptional warmth and a funny sense of humor. And despite that, he wasn't a bad pool player either. So, we miss Steve. And it's just a tribute to him to how many people he's impacted.

With that, then the next order of business is some clarifications. There were some conversations that were off record yesterday that I'd

like to have put on the record. If we could start with Susan Nagel and Dr. Paul Friedman.

DR. SUSAN NAGEL: Let's see. I had a question. I'm not sure what slide number it was of Dr. Paul Friedman's, but it's slide number 97 of the PDF that was sent to us; and that is in looking at the number of chemicals that impacted one, two, three to ten hormones.

My question is there's a blue box around the number of chemicals that impacted three or more hormones. And so, I think my question could be summarized as how many chemicals are in that box versus in the title of the slide it said 656 chemicals were selected to move forward. There's like -- and I guess I can -- I'm not sure if you know the answer to that question.

The first question just being how many chemicals are inside that blue box.

DR. KATIE PAUL FRIEDMAN: Okay. This is Dr. Katie Paul Friedman. Your question this morning, as I understand it, is you just would like to know how many chemicals are in the blue box here. And that's a number approaching 500, just approximate. We

screened 656 chemicals in concentration response out of 2012 screened in single concentration.

But there's always little caveats to that, right. Predominantly we moved chemicals from single concentration to multi-concentration screening, based on the criteria that the chemical perturbed at least three steroid hormones in the panel. But we also advanced almost 150 chemicals for which that was not true.

We also advanced a number of chemicals that were interesting as reference chemicals or chemicals of classes of interest of the Agency, like phthalates that maybe did not satisfy that criteria in single concentration screening.

You know, extending your question and thinking about it the way that we discussed a little bit yesterday, moving ahead to that Venn diagram that I showed.

DR. SUSAN NAGEL: Yes, you're absolutely dead on. The question evolved quickly, but I was very interested in -- which I think she's going to speak to with the Venn diagram though in that slide of 97 -- of the chemicals that have impacted one or two hormones, how many of those, since there's a

difference. There's 500-ish that are in the blue box, and there's 650 that moved forward. There's 150 other ones which you just spoke to a little bit. I was just curious, then, how many of the 1 or 2 that moved forward.

DR. KATIE PAUL FRIEDMAN: Right. I unfortunately don't have an exact number for you, but, you know, approximately 150 chemicals moved forward that came from the 0, 1, or 2 bars. They're distributed across there.

And to give an idea of why perhaps we maybe prioritize this way, you know, thinking about prioritizing chemicals that maybe have the most impact on steroidogenesis in this particular H295R model, you'd probably want to look at the chemicals that seem to perturb more hormones. And I'm thinking practically with the parameters that are imposed on all aspects of science, right, is money.

Just giving approximate figures, if there are 400 chemicals that perturbed only one or two hormones, to screen those in multi-concentration response would have been well over a million dollars. This resulted in a great savings to the program and

allowed us to screen chemicals that we hypothesized had a greater impact on the system.

DR. SUSAN NAGEL: And so then as a follow-up question to that, of those chemicals that there's a hundred and -- well, we don't know. The 150-ish chemicals that came from 0, 1, or 2 columns, how many of those chemicals, then, did you compare the results of this initial with the full concentration response? How many of those chemicals actually -- they actually did go on to impact more than 0, 1, or 2 hormones?

DR. KATIE PAUL FRIEDMAN: You mean for some Chemical X, if in single concentration it only impacted one hormone, but then we actually happen to screen it in multi-concentration and perhaps it impacted 2 or 0?

DR. SUSAN NAGEL: Or more. Yeah.

DR. KATIE PAUL FRIEDMAN: Right. We in our initial work, the Karmaus et al., we did look at recall sensitivity. And in that work, there were about 120 chemicals, that were advanced, that only perturbed zero to three steroid hormones. And if you lumped those together, the recall sensitivity was

close to 50 percent as Dr. Pullen Fedinick mentioned yesterday.

The recall, the ability to create a concordant result in a second screening was not as good for chemicals that only affected a few hormones. But we haven't repeated the analysis since running even more chemicals to look at the specific recall.

DR. SUSAN NAGEL: But you have those data that you could look at? Because then you moved those chemicals forward and then you tested them; so you just haven't done that particular analysis for this?

DR. KATIE PAUL FRIEDMAN: There may be a very small number of chemicals that we could add to that. Subsequent to that Karmaus et al. work, we went back and screened chemicals that perturbed only three steroid hormones. In that initial work, we had had a threshold of four.

So, there's an iterative screening process where we screened additional blocks of chemicals. And so, I just haven't looked back to see how if we recalculated recall, let's say there were 120 chemicals, that only affected zero to three steroid hormones. I haven't redone it now that we've

screened about a hundred chemicals more in total. So, the numbers might revise slightly.

But, you know, keep in mind also that that Karmaus et al. work, while completely valid, uses a ToxCast Data pipeline which we didn't use in the analysis presented yesterday and today. You're looking at different types of analysis. The numbers might shift depending on how you do the analysis and your hit call-in.

DR. SUSAN NAGEL: Thank you.

DR. KATIE PAUL FRIEDMAN: I wanted to add just another clarification based on what we discussed yesterday afternoon. We had brought up this Venn diagram, and I think it's important to just reiterate. I think everyone saw this, but here in the Venn diagram, we only have four ellipses. This is about steroid hormone classes and not the number of steroid hormones hit.

For example, this 1 chemical that hits only the androgen steroid hormone class, it's possible it hit one hormone, but it's possible it hit two. Or this one chemical that hit only estrogen and androgen steroid hormone classes may actually have hit four steroid hormones because there are two androgens and

two estrogens. I just wanted to clarify that aspect of your question.

And, you know, the other clarification we talked about related to the screening implementation versus, you know, what we found in multi-concentration screening and concerns that maybe we were missing things by not screening every chemical and multi-concentration. You know, I'd like to maybe frame that issue a little bit with respect to the charge question.

You know, there's a screening implementation that we executed in order to screen as many chemicals as possible, given our parameters, through multi-concentration screening. But if you were looking at this as an alternative for the EDSP program and for just estrogen and testosterone and replacing the low throughput H295R, then I imagine you would have chemicals of priority and you would screen those in multi-concentration.

And so, in terms of using this as an alternative, I imagine that you would be looking at using the multi-concentration version, not just a screening implementation that we ran in an effort to try to screen as many of the relevant chemicals in our

list of over 2,000, right. You would have a regulatory imperative there to do a concentration response screening, so.

DR. SUSAN NAGEL: Yeah. I think that's a super important clarification. And it's a little confusing in the charge questions of presenting the data -- you know, the work that you did versus how would it be implemented. So, great.

And then, yeah, I think as far as the Venn diagram goes, yes, those are groups of hormones that were impacted; androgens being however many are in there. And then I think the question, though, really that I had has already been addressed, about that.

DR. KATIE PAUL FRIEDMAN: Great. Thank you.

DR. JAMES MCMANAMAN: Dr. Bever, did you have a comment about this, too?

DR. RONNIE JOE BEVER: Yes, I did, but Dr. Paul Friedman pretty much covered it. I tried to, yesterday, specify that as the Endocrine Disruption Screening Program would mandate when we did our testing call-in, that it would be multi-concentration. So it's clear-cut. It will be multi-concentration.

As Dr. Paul Friedman said, the single concentration thing was simply a way to efficiently use the resources. It's not going to be the way that we carry out the data call-ins. Those will be multi-concentration. Thank you.

DR. JAMES MCMANAMAN: Thank you. The other conversation that we want to bring on the record is between Dr. Clewell and Dr. Schappelle. If you could tell us what the question was and what the response was.

DR. REBECCA CLEWELL: Okay. Hearing this conversation, now I'm getting two -- I think the impetus for this conversation was along a similar vein as that conversation. I had asked a question about the charge question to the whole panel. And I'm already having trouble. I'm sorry. I was thinking about this so hard last night.

I was wondering about the charge question and the difference between the sort of very exact nature of the charge question for the steroidogenesis assay in which the question is, can this low throughput -- or the high throughput directly replace the low throughput? And also, the same sort of question for the androgen receptor. Can these

eleven assays in the androgen receptor model directly replace the low throughput androgen receptor-binding assay?

That's a very specific question. And it feels different and causes a little bit confusion in my mind, and I believe by listening to the rest of the panelists, maybe theirs as well; because the introductory slide with all of the different -- where it shows the potential replacements of the current EDSP assay, shows the AR lit up and highlighted and many different replacement assays.

It also shows the ER in the steroidogenesis -- is there any way we could bring that up because I feel like I'm not going to describe this well.

And so, even though it says currently right now we're trying to replace the AR binding assay, so that's proposed here, and it's highlighted as proposed. And only for the steroidogenesis model. So, we're doing a one-to-one replacement.

And then there's in parenthesis, "future" for the in vivo test. It's shown here in a highlight, and it's introduced in that way. And so, there's conversations happening here, I believe,

amongst like yesterday what was happening publicly, that made me think that some other folks might be having a little bit of cognitive dissonance about this -- as I was -- in terms of the charge question is very specific, but the goal of this is much broader.

And so, while the AR model right now may be clearly replaceable, the concerns about things like off-target nongenomic effects and other concerns like that -- not off-target, I'm sorry, but nongenomic effects is more a concern when the AR is moving to replace a whole animal in vivo, right. That's a different question, but it's not entirely separateable from the discussion of the AR assay model.

I don't know if I'm explaining this very well today. But that was my question. That was my underlying question when I asked about the charge question. And Seema came to kind of clarify about the goal of this particular meeting versus the goal of the overall program in general and it was moved towards in vitro.

DR. SEEMA SCHAPPELLE: Yeah. Thanks, Dr. Clewell. I think you've actually done my job for me where you've asked the question and you're

answering it as well. But let me further clarify just a little bit.

The intention of the alternatives that are being proposed within the program, from a broad perspective, are to achieve alternatives as coverage for the full Tier 1 battery as we progress over time, but we're not there yet. That's going to be a process that's going to take some time. We are here today, in this meeting, focusing on the one-to-one replacement of the AR model for the androgen receptor-binding assay, and similarly, for the steroidogenesis approach as well.

Yes, we're segmenting. We're doing this in steps as we progress, and that's the portion that we're focusing on right now.

DR. REBECCA CLEWELL: Okay.

DR. JAMES MCMANAMAN: Thank you. As a reminder, this brings up the point is that we're here to evaluate the charge questions and the science behind those charge questions as they specifically apply. And while it's important to understand the overall goal, you know, it's like any scientific question is, that we have to evaluate it as it's stated. And if it's not stated clearly or if there's

a gap in there, then that should be brought up in the discussion of the charge question.

Let's see. That ends the two conversations. And it's fine to have the conversations. It's just that we need to put them on the record so that it's available to the public.

At this stage, I think we're at follow-up questions from the previous day. And so I'll open it up to the panel if there are other questions that we need clarification for. We can do that right now if you have it.

Yes, Dr. Pullen Fedinick.

DR. KRISTI PULLEN FEDINICK: This is Kristi Pullen Fedinick at NRDC. It's any question and all that we may have had that have come up or what was the –

DR. JAMES MCMANAMAN: Well, it should be related to the topics, but.

DR. KRISTI PULLEN FEDINICK: Yeah, I was curious what everyone's favorite color was and, you know, what music you listen to when you're getting up in the morning.

DR. JAMES MCMANAMAN: Yeah. So anything that you need clarification on.

DR. KRISTI PULLEN FEDINICK: That's great. I actually had a question about the steroidogenesis assays. This would apply for the EN, so charge questions, I guess, 2 and 3. And so I was curious whether or not the Agency had done tests to look at how the results compared to the Tier 1 List 1 results.

The OECD comparisons that you've shown so far are really just for standardization, right, for the validation process for that. But, did you look at whether or not you were able to capture the Tier 1 List 1 H295R assay results?

DR. SEEMA SCHAPPELLE: I'm going to ask Kristan Markey to come up as well as Katie Paul Friedman. If you guys both can provide your perspectives, that would be great.

DR. KATIE PAUL FRIEDMAN: That's a great question, Dr. Pullen Fedinick. There were a number of chemicals from List 1 that did have H295R assay results. But keep in mind that those results were interpreted by the program offices as part of a weight of evidence. And so, the determinations made were really integrated within that weight of evidence analysis.

And so, our exact hit call-in, based on single assay results, are not necessarily comparable because they had a lot more data to look at for the entire data package that was submitted for that chemical. Because of the different purposes of those experiments, they're difficult to compare and not necessarily a fair comparison, would be my opinion.

DR. KRISTI PULLEN FEDINICK: Could I just comment? Wasn't that done for the AR model, so we looked at the results of the List 1, Tier 1 assays and compared that? Could we also not -- I mean it seems as though we should be able to do that scientifically for steroidogenesis alone as well.

DR. KRISTAN MARKEY: That work is ongoing.

DR. KRISTI PULLEN FEDINICK: So then right now we can't compare it to how it's actually worked in practice outside of just the validation efforts?

DR. KRISTAN MARKEY: It hasn't been done. I mean you are free -- the data is publicly available, but it has not been completed yet in time for this SAP meeting.

DR. JAMES MCMANAMAN: Other questions? Sure. Dr. Nagel?

DR. SUSAN NAGEL: I'm very curious about your comment yesterday, Dr. Paul Friedman, about the 70 percent cell viability as measured by the MTT assay. And so, you said that was within the realm of the variation of the assay. I'm curious how that was determined. Was that the standard deviation? Was that the CV? Was that an LOD, an LOQ? Just because I am very concerned with using that number, but.

DR. KATIE PAUL FRIEDMAN: I have a slide about this.

DR. SUSAN NAGEL: Awesome.

DR. KATIE PAUL FRIEDMAN: I'm pulling it up right now. While they are working on that, just to give you some background on the ToxCast program, typically in the ToxCast Data pipeline, which was used to analyze the MTT data, we approximate what's called the baseline median absolute deviation. And basically, that is the median absolute deviation around a baseline that's defined by the user of the pipeline.

The default on that is the activity at the two lowest concentrations screened. For most of our assays, there's very little activity across the two lowest concentrations, and it gives us a conservative estimate of maybe a larger estimate of what the noise is around baseline. And then, typically, we use a multiple of that baseline median absolute deviation to create a threshold for hit calling.

As an example, typically we might use three times the baseline median absolute deviation or five or six to delineate between something that might look like background noise, versus something that's truly a signal in the high throughput screening assay. And similarly, we've done that for the MTT data. Let me show you what that looks like.

Okay. I'm bringing up the slide. As you suggested, we did use a 70 percent cutoff for cell viability as determined using MTT data as an indicator of cell viability. But most of the MTC data corresponded to cell viability of greater than 80 percent. What that means is that for the top concentration used for any given chemical, this is a distribution of those viability data.

And you can see here on the X-axis the percent cell viability versus the frequency. And almost the entire distribution falls between greater than 80 to a little over a hundred percent because we don't constrain our curves to a hundred percent. You can see only 35 out of the 671 samples -- this is not unique chemicals -- in multi-concentration, had a viability between 70 percent and 80 percent. We're not talking about actually a large fraction of the library falling into that zone.

And then here annotated in the dashed red lines is actually five times the baseline median absolute deviation for the MTT assay, as determined using the ToxCast Data pipeline. And so you can see that actually that bounds the data in this distribution quite well. You can see a little bit of information that's just slightly outside of it, but those bounds tend to encompass the distribution.

We think that probably looking at effects that are 10 to 20 percent cell viability in this assay, it's likely that that's within the noise of the baseline based on the way that we're running the pipeline.

DR. SUSAN NAGEL: But that was determined on the two lowest concentrations of test chemicals as opposed to the variation of the vehicle control?

DR. KATIE PAUL FRIEDMAN: That's correct.

DR. SUSAN NAGEL: What is the rationale for that?

DR. KATIE PAUL FRIEDMAN: It's a conservative estimate of the background activity. Typically, in the two lowest concentration screened for most of our assays in the list of assays that we have, there's very little activity, and it gives you more samples from which to pool across the plate.

A lot of times ToxCast Data, we're including data from vendors that are designing plates and they might always put DMSO in the top-left corner of a plate. Anyone in the room can think about why there's so many reasons not to do that. This gives us a better sampling across the plate and really a better indication than just using the DMSO.

And we do have some assays and vendors that will randomize their plate design. And in that case, you might find it more acceptable to use DMSO,

but this was not the case with the H295R MTT data. They were not randomized plate designs.

DR. SUSAN NAGEL: I guess that's a whole other thing.

DR. KATIE PAUL FRIEDMAN: Yes.

DR. SUSAN NAGEL: That doesn't seem completely unacceptable.

DR. KATIE PAUL FRIEDMAN: How do people screen data analysis and platelet out, it's a lot of experimental details. But we think that using the two lowest concentrations, typically, is giving us a better consideration of what the variability around the baseline is.

But in some cases, we might choose DMSO if we can show that it truly is representing the baseline variability in the assay, and that's just a matter of proving that.

DR. SUSAN NAGEL: Yeah. I mean, I like the rationale. I do not like using the two lowest concentrations. Just because, by definition, they will increase the variation. But I mean, you feel comfortable with that as far as being conservative?

DR. KATIE PAUL FRIEDMAN: We're typically not seeing activity. For instance, if you

remember any of the plots yesterday, the example plots, for the mean Mahalanobis distance by concentration. Almost all of our mean Mahalanobis distances that exceed the critical limit are the top three concentrations and not the lower two. It's kind of the exception to see that.

For most chemicals, the lowest two concentrations, in this assay and in many other assays, are not active. And if they are, again, it's just giving you a conservative estimate of that variability. But that's just a data processing choice that can be changed.

DR. TODD PETERSON: This is Todd Peterson. I just want to add that the supplemental slides that are being shown, during questions and answers, I will be providing them to the panel and then after the meeting, I'll upload them to the docket so the public can have access to them as well.

DR. JAMES MCMANAMAN: That question was from Dr. Nagel and Dr. Paul Friedman answered it. This is Dr. McManaman. I have a follow-up question about that. Is this data -- or these data, rather, are they from your laboratories or are they a composite from all of the vendor laboratories?

DR. KATIE PAUL FRIEDMAN: For the data presented here, the MTT data were generated initially by CeeTox who was then bought by Cyprotex who generated it. And same for the H295R cell culture. CeeTox was not equipped to do the HPLC-tandem mass spec measurement of the steroid hormones, and they subcontract that to Opands (phonetic) who is a company that basically just that. It was a collaboration between those two vendors to deliver the data set.

DR. JAMES MCMANAMAN: Yeah. I guess my question was whether the variability that you're seeing is due to the inherent variability of the assay or the variability of different providers giving -- because, like you said, some put up in the top left-hand corner, some randomized it. The question is about whether the variability is due to where the assay was performed or whether there's really inherent variability.

DR. KATIE PAUL FRIEDMAN: All of the MTT data were generated by CeeTox/Cyprotex within the same group.

DR. JAMES MCMANAMAN: Thank you. Dr. Clewell?

DR. REBECCA CLEWELL: I think we need to be careful about how we use the word conservative, especially when we're in the realm of risk assessment, even though we're not talking about the risk assessment today. Conservative has a lot of connotation to it.

When you're looking for a hit in an assay in these high throughput screening, then expanding the range of variability so that it makes it harder to get a hit is conservative in the way that it makes it harder to get a hit so you're less likely to get a false positive.

Here, what's happening is we're using the viability – we, you all -- are using the viability to judge where a hit is viable or reasonable or not. And that's actually having the opposite effect, right? So, instead of making it harder to get a hit, because this is an inhibition assay where it's based -- it's steroidogenesis based on mitochondrial function, then spreading the range and making it more permissive to have unhealthy cells in your assay, is actually making it more likely that you will get a hit.

These are opposite effects. And I just want to be careful because conservative has a

connotation. In this case, you actually could make the argument that this is conservative for risk assessment because you're more likely to get a hit.

DR. JAMES MCMANAMAN: Dr. Clewell, this sounds like a discussion for the charge question.

DR. REBECCA CLEWELL: Sorry.

DR. JAMES MCMANAMAN: It's not --

DR. REBECCA CLEWELL: Right.

DR. JAMES MCMANAMAN: We want clarification.

DR. REBECCA CLEWELL: Okay.

DR. JAMES MCMANAMAN: But it's an important point, but we should probably bring it up during the charges.

DR. REBECCA CLEWELL: Okay.

DR. JAMES MCMANAMAN: Okay?

DR. REBECCA CLEWELL: I can do that. But I don't know if she wants to respond to that.

DR. KATIE PAUL FRIEDMAN: Can I just respond to that?

DR. REBECCA CLEWELL: Yeah.

DR. KATIE PAUL FRIEDMAN: Yeah. When I say conservative estimate of the baseline, I just mean that we're including more noise in our estimate of the

baseline, no illusion to risk assessment or anything like that.

And you're correct. In relationship to the OECD test guidelines, which allows 80 percent viability, we allowed 70 percent. And in this case, that's only 35 out of 671 samples that fall within that 10 percent range. But it basically allowed us to look at more samples, and more concentrations of that sample, to look for a response on steroid hormone biosynthesis, which was really the objective.

In addition to looking at how our baseline noise looked, it's also, I think, helpful because it allows us to include perhaps an additional concentration to see if there's concentration response behavior that we can interpret. It just gives us a little more information. But, again, it impacted extremely few samples, but it seemed to be a very defenseable baseline just looking at our data and our interpretation of the data.

DR. JAMES MCMANAMAN: Other questions? I guess it just goes to show you that in DC nothing conservative can be taken simply.

Okay. I think we can move on then to the public commenters. I don't know exactly how you

guys are going to do this. Ellen Mihaich, Rick Becker, Steve Levine, and Brandy Riffle. There may be some substitutions I understand. I understand as a group, you have 50 minutes; is that right?

DR. ELLEN MIHAICH: Probably approximately that, yes.

DR. JAMES MCMANAMAN: Okay.

DR. ELLEN MIHAICH: Okay. We'll keep going as fast as we can. That's why we're all here so we can just move right down the line.

DR. TODD PETERSON: And be sure to announce yourself --

DR. ELLEN MIHAICH: We will.

DR. TODD PETERSON: -- and affiliation before speaking. Thank you.

DR. ELLEN MIHAICH: Yeah. Thanks. Hi, my name's Ellen Mihaich, and on behalf of the Endocrine Policy Forum, I want to thank you for allowing us to speak today. I'm just going to present a few introductory comments.

As I said, my name's Ellen Mihaich. I am an ecotoxicologist, risk assessor. I am the owner and principal scientist of Environmental and Regulatory Resources, a small consulting company in

Durham, North Carolina. I am also an adjunct professor at Duke University, and I teach risk assessment, and the scientific coordinator for the Endocrine Policy Forum.

The Endocrine Policy Forum is a consortia of List 1, primarily List 1 test order recipients and other stakeholders. We are a self-funded group and we represent more than 95 percent of the people that got test orders and had to do all of the initial screening for Tier 1. We have a lot of experience within our group and a lot of experience with the Tier 1 assays and evaluating them.

We also have additional stakeholders, as I have shown on this slide, like CropLife America, American Chemistry Council, the American Cleaning Institute, Consumer Specialty Products Association, the American Petroleum Institute, and some various consulting companies.

Our main objective as we went into this whole program was to be able to better understand how to perform and evaluate the screens that were in Tier 1. So, we were addressing technical guidance and science advocacy, and focusing ultimately on a very balanced risk-based analysis in order to properly

screen, test, and regulate chemicals for endocrine activity and the potential to cause endocrine adverse health effects through the endocrine system.

And it's also becoming much more important that we do this because around the world, there are different regulations, and we would like to be as harmonized as possible to better evaluate our chemicals.

And to us, the dialogue is very important, so that's why we are very involved, and we have been from the start in providing comments here; as we evaluate things, listening to how you guys evaluate it. Because it's very important to listen to the different expertise come to the table to best inform this kind of testing. We have provided, as I said, public comments at all of the EPA SAPs. We're kind of like groupies I guess.

We present platforms, posters. I've been teaching, with the help of many of my colleagues, short courses, at least at two SETAC meetings a year on the endocrine system, and not only just -- you know, and the screens and the tests, and the regulations, and had to evaluate it from a risk-based approach.

We also have done quite a bit of publishing. We started back in 2011, with a series of publications on developing a weight of evidence process to evaluate these screens and tests. And that was very important because we quickly recognized that we're going to have a lot of new data. I mean, there's approximately 89 endpoints in the original Tier 1 set of studies, the 11 set of studies. Trying to put all those together from a hypothesis-based perspective was going to be very important, and we did that.

Because we're a consortium of the people that have done this testing, the other thing that we were able to do in 2015, was actually get a paper together by Adam Schapaugh, et al. on looking at just the normal control variability. We have all of these 11 screens and tests, and there was a lot of issues with failing when we were doing these because of not meeting the acceptance criteria and then having to repeat them. And so it was, I think, very worthwhile to come together and share the data to look at just what's the normal control variability.

And then just recently we've moved on because there's been such a discussion about potency

and threshold. And my colleague Chris Borgert just had a paper accepted with revision in Archives of Toxicology looking at the potency threshold ER alpha agonist, and trying to set a threshold to say and to show that at some point there's just not enough activity to cause an effect.

We do fully support the development and use of high-throughput methods, and we really do commend the EPA on the amount of work and time that they've put into this, and care in making sure that these things are effectively vetted. And we do think ultimately it will be able to help them fulfill the statutory mandate to screen for potential endocrine activity, as well as reduce animal testing.

However, we really continue to push that scientific confidence in this methodology needs to continue to be established. We need to know that there's comparative responses, and I think we've seen a lot of that in the discussions that we've already had from yesterday. Being able to look at intra- and inter-lab repeatability/orthogonal assay comparison.

One thing that's very important for us, is if we're going to have to use these things, these methods, they need to be transportable and usable and

optimized. Just in the EDSP-21, which are the more endocrine-focused assays, there is, I think, four or five difference vendors. It's very difficult to then go to each one of these vendors to get these things done. EPA did a nice job of discussing optimization for the ER. We're anticipating you'll the same thing ultimately for AR. And so, it's a good way forward, but we just continue to be interested in this.

And they need to be fit for purpose, and we need to understand what that purpose is. And that's important because activity is not disruption. And, I think, that's something that unfortunately people lose sight of sometimes. And so, these Tier 1 screens or bioactivity measures can't do more than identify the potential for endocrine activity at this time. And so they don't identify adverse effects. There is more that will go on after this, so, just to keep that in mind when you're evaluating these.

And that's important to also consider the fact that there is a definition of endocrine disruption that is pretty globally accepted, and that's from the World Health Organization IPCS in 2002, where an endocrine disruptor is defined as, "an exogenous substance that alters function of the

endocrine system and consequently causes adverse effects in an intact organism."

What that says is that you have to have a causal link between the mechanism and the adverse effect to call something an endocrine disruptor. Otherwise, it's simply endocrine activity. And plausible effect is not good enough, or plausible link. Causal link needs to be what it is. And I think what we're doing here is talking about being able to look at the mechanism and then ultimately link that, at some point, with that adverse effect. And I think that's a really robust way to do it.

With that, I am going to move on to our next speaker, Chris Borgert. And I'll let him introduce himself.

DR. CHRISTOPHER BORGERT: Okay. Thank you. And, again, I appreciate the opportunity to speak here. I am the replacement speaker, so I'll be presenting slides that were originally prepared by Rick Becker. And I'm a bit handicapped in that regard, but I'll do my very best.

First of all, we do strongly support the pivot from the initial EDSP Tier 1 battery to these more advanced molecular and cellular screening

assays. It's a much more efficient way to generate mechanistic level data. And we do commend EPA in not only undertaking this effort, but making the assay information, the results, the models, the code, et cetera, publicly available; that enables independent evaluation and it enables independent use ultimately.

And I just want to pick up on a point that Dr. Mihaich made, is that going to five different vendors is much easier in a very, very large scale, as EPA has been undertaking. But when individual companies, for example, have smaller batches of chemicals, that become a much more onerous kind of effort and more expensive. And I'll get to that in a minute.

We also believe that it's very important that this transparency be continued and that the reporting continue for the other models -- we have every confidence EPA will do that -- because we need to establish confidence in these methods if everyone's going to continue to use them, and improve and enhance the use of them.

We have a recommendation in terms of improving and enhancing transparency for the new methods. And we would recommend the Scientific

Confidence Framework that was published by Cox, et al. in 2015. It has some guidelines for specifying things such as analytical data, the replicability of the techniques, the applicability of the domain, performance metrics of the prediction models, those sorts of things. And so we want to recommend that that process and that method for transparency.

Now moving on to this issue of the utility outside of EPA's program. Dr. Mihaich mentioned that we're delighted to see that EPA optimized the ER expert model. And I think there's a publication that if it's not out already, I know it's accepted for publication. It's been presented by EPA. Where that original set of 18 assays was compared against various subsets, and it was found that there were some subsets as small as four assays that were actually highly predictive of the overall 18-assay model.

We would encourage that that be done as soon as possible with the AR model. And I don't know if you would call it optimization or, you know, maximized efficiency. I'm not sure. We're using the term "optimized", but what we're talking about there is the greatest predictive value for the fewest number

of assays, because that just enhances the ability of others to use the system and generate more data on more chemicals.

We think that should be, if possible, an integral part of the effort. I'm not familiar with all of the techniques that EPA is using, but if it's possible, to conduct that optimization while the larger model is being constructed and at least doing it as soon after the larger model is developed, we would encourage that.

All right. I'm going to get now to some of the issues around potency. And EPA indicates there's an intent to match up bioactivity in the AR model, as they did in the ER model, with exposure for purposes of priority setting. The in vitro/in vivo comparison to convert these is important. And so there are some problems that arise from the use of these AUC curves because -- and I understand why EPA does this. The AUC curve is integrating the results of a number of different assays and EPA has well described that. But it does skew, in some regards, the relationship to potency for any one of the assays or any set of assays for a particular modality, for instance, ER alpha or ER beta.

And so if you're going to actually look at the strength of activity, in other words, the potency of a chemical and compare that to the potential exposure, it's important to have a somewhat accurate reflection of that potency. And while I understand that that's very difficult to do across a set of even 11 assays -- in fact, impossible to do -- it's important that that information not be lost.

Currently, the AUC scores have to be reconverted. I'll go on to the next slide. Now, I think EPA answered -- these were provided in our written comments. And I didn't prepare these analyses, so I'm not really prepared to speak to exactly what was done, but I'll speak to the larger point.

I think EPA has described yesterday what the relationship is between the activity and the AUC curves. But I want to recall an example from the ER model, and then make the analogy to why the potency data is important. When EPA discussed, in their review document, the ER model, there was a false negative noted; and there is some discussion about the potency of that false negative.

And I've presented to this group -- I know the membership has changed somewhat. But I presented regarding the potency of that particular chemical, which was 10 to the minus 6 as potent as 17 beta estradiol. It was even less potent than some of the non-aromatizable androgens.

And so, in the current EDPS-21 dashboard, that chemical is a negative, but it was important to go back and understand the potency data of that chemical, especially -- and you can verify that it, in fact, doesn't produce adverse effects by an estrogenic mode of action. That kind of information, I think, needs to be systematically sought whenever the AUC data are interpreted, especially for some of the weaker chemicals.

We just want to encourage EPA to have some systematic way of doing that. Perhaps, they've implemented that already, but we think that's useful information that can come out of this but isn't immediately obvious from the AUC curves.

Then we can go to the next slide. And I'm going to gloss over these because I think I've already made the point here, and EPA has explained

some of the relationship. So, you can go to the next slide.

I think I've made the point that understanding that relationship between the AC50 values and the AUC scores is important. Right now those details are not fully available, I don't believe; but to the extent that you can make those more clear and more available, that would be helpful and encouraged.

You can go to the next slide. Again, I'm going to pass over that slide and go to the next one. And pass over that one because I think EPA explained some of this yesterday.

Then the conclusions of my remarks are that we strongly support the pivot to this high-throughput model. We think that independent replication of these models is necessary, and so we commend EPA's efforts to make all of this transparent. I've discussed the relationship between the AUC and AC50 scores.

We would encourage some systematic use of the potency data that's available from the individual assays, in addition to the way the Agency is using the AUC scores. I recommended a Scientific

Confidence Framework as a core element. And we would strongly, then, support ultimately integrating these with exposure.

That concludes my remarks. And, again, I thank you for your attention.

DR. STEVEN LEVINE: Good morning, everyone. My name is Steve Levine. I'm a senior science fellow with the Monsanto Company. And I'm our environmental assessment strategy lead, but for about a decade I led our global eco-toxicology and risk assessment function.

And I'm going to be giving these comments on behalf of the Endocrine Policy Forum. And I wanted to thank the DFO and the panel for the opportunity to provide these comments on Charge Questions 2, 3, and 4. These are all on the H295R steroidogenesis assay, and I think these are going to generate some continued discussion on that assay.

I wanted to mention, I feel relatively close to this assay in that I had the opportunity to serve on the EDMVAC from 2004 to 2006. That was an advisory panel that was put together by EPA, after the EDSTAC, to help advise EPA on the development and

implementation of the Tier 1 batteries and eventually the Tier 2 tests.

And at our very first meeting, we made the move away from the mens' testes assay, which was the original assay looking at steroidogenesis. There were some pretty significant issues there, and EPA had recommended moving to the H295R. And we had some recommendations, at that time, which I'll come back to through the course of my talk.

Here's just a quick overview of what I'm going to cover. I need to cover a little bit about proper dose setting for endocrine screening assays. And this is going to focus on in vitro [sic] but these same comments really apply for in vitro as well.

I want to spend some time talking about the basic biochemistry of steroidogenesis, because I think that's going to serve as a good foundation for some of the comments I have and will complement what we heard yesterday. Then I want to quickly just go over a summary of the results, and spend a little bit of time on a case study to drive some points home, and then close with some recommendations.

We had a lot of discussion yesterday on the outcomes of in vitro assays. And those can be either specific or non-specific, with specific or direct effects being biomolecular interactions against targets, such as receptors and enzymes or pathways, that occur below concentrations that disrupt normal cellular processes or cellular machinery.

Non-specific, or really indirect, effects are interactions that can occur at levels that disrupt normal cellular processes or cellular machinery. And, again, that can occur in cell free systems or cell based systems. An example of in cell free systems could be pH effects, protein denaturation, such as receptors, changes to protein and protein interactions. An example of that maybe talking -- the coactivator recruitment assay we were discussing. And disruption and enhancement of binding kinetics.

And I think the one I wanted to just say a few words about is protein denaturation, because that came up as an issue with the estrogen receptor/ competitive receptor-binding assay. The Endocrine Program had requirements to test up to 1000 micromolar. That's a relatively high concentration.

And that caused some issues with the validation work for the estrogen receptor assay, and a lot of false positives. So, EPA -- Susan Laws' group actually went back and did some secondary analyses; some secondary plots to really help distinguish between direct and indirect effects on that competitive receptor binding assay. And I think it's worth going back and looking at that paper because it had some very good approaches to really distinguish between direct and indirect effects.

For cell based systems, cytotoxicity or effects on the cell can be the result of DNA or lipid reactivity, disruption of proteins or cell membranes. The one I've highlighted here is mitochondrial disruption. That's what I'm going to spend some time talking about. And that's important because that's where steroidogenesis initiates.

We also can see oxidative stress, apoptosis. And I've also highlighted stress to the endoplasmic reticulum. That's important because once the first steps of steroidogenesis take place in the mitochondria, progesterone -- excuse me -- pregnenolone goes out to the cytosol where additional biotransformations occur. Those P450s are anchored in

the smooth endoplasmic reticulum where they're associated with reductases, which supply them with reducing equivalence to catalyzed or biotransformations.

Impacts to the smooth endoplasmic reticulum could either increase or decrease enzymatic activity. And those of us that have done P450 purifications really understand that; because you can have artifacts of P450 activity, either from increases or decreases in activity.

Because of the potential for non-specific effects, dose setting really takes on greater importance for specific endocrine MoAs, okay. We're not simply testing for a tipping point in normal cellular function with these H295R cells. Rather, we're looking for a specific effect on steroidogenesis.

In this endocrine battery, we're testing specific hypotheses. We're testing for an adverse effect through an endocrine mechanism, not simply an adverse effect.

Here's just a quick overview of the basic steps of steroidogenesis. And, as I said, this is going to be important as a foundation for some of

the comments that I have to come. Steroidogenesis begins when cholesterol is shuttled from the cytosol to the mitochondria, and that's done by the StAR protein, or the steroidogenesis acute regulatory protein.

And then cholesterol comes through the mitochondrial membrane; so it goes from the outer to the inner mitochondrial membrane where the biotransformations can start. What I have highlighted here in red, is it's essential for the mitochondrial electrochemical gradient to be functional for this to happen. If that gradient gets shut down, steroidogenesis gets shut down. The StAR protein cannot bring cholesterol through that mitochondrial membrane.

When the mitochondrial membrane is functioning, the StAR protein is internalized. It's a 37 kilodalton protein. It comes into the mitochondria. It's cleaved and inactivated to a 32 kilodalton protein. That's the inactivation step.

And then you go from pregnenolone to progesterone, then out to the cytosol where the additional steps in biotransformation, or steroid synthesis, take place. And, again, there can be

stress to the smooth endoplasmic reticulum. That's why the cytotoxicity assessment is so important.

I wanted to give a quick overview of the result. And my first impression, and after working in this field for years, is it was a relatively high hit rate. Out of the approximately 2,000 chemicals that were tested, 600 odd chemicals had concentration responses. That's a third of the chemicals. That's a high hit rate. Yes, there were a number of compounds in the library that had specific modes of action for inhibition of steroidogenesis. You know, we heard about the triazoles, the imidazoles, that's a well-understood mode of non-competitive inhibition.

We saw positive responses for greater than or equal to four hormones for 500 chemicals. We had a summary of that in that blue box. Those were the ones that went into concentration response. From the Venn diagram, we could also pull out that about 300 chemicals, or about 15 percent of the chemicals, were positive for at least one hormone in four hormone classes, okay. So, that's androgens, estrogens, progesterone, glucocorticoids.

To me, initially, that looks like a non-specific response. It's hitting something very early in the pathway; and that could be mitochondrial toxicity. But that was a stand out for me.

The high hit rate could be an outcome of cell stress and an effect of mitochondrial function. We saw the limit of greater than or equal to 30 percent cytotoxicity. And I thought maybe initially that was the reason. And after looking at the data more closely, and particularly the presentation, the supplementary slide that was presented, I don't believe that's the case.

And 30 percent is what I'll call the LOD, the limit of detection, for that assay, and that was explained well by the standard deviation assessment. But it could also be related to an insensitive cytotoxicity assay. That MTT may not be the right assay for this H295R steroidogenesis assay. And the reason why I'm saying that is, yes, the MTT does look at mitochondrial function, but only about 50 percent of the activity comes from the mitochondria. The other half comes from NADPH oxidoreductase as in the cytosol. So, it has some specificity, but it may not be the right cytotoxicity assay.

And the white paper recommended that markers of mitochondrial toxicity need to be incorporated in a systematic and quantitative matter. And that's called out on pages 106 and 109. And I think that's an important point. The other point that was raised, was perhaps in the future they could look at movement of cholesterol from the cytosol to the mitochondria, okay.

That's not a trivial thing to do. You can look for the protein in the cytosol. You can also look for it in the mitochondria. If you see it accumulating as a 32-kilodalton protein, you know steroidogenesis is functional. Again, you can do it. I've done it, but it's not trivial.

This is an example of an effective steroidogenesis assay. And this is from a paper I published back in 2007, and had done the work several years before that. We're working with the JC-1 assay very early in its development, and this is an assay that looked specifically at the electro chemical potential in the mitochondria. And I was working with Vassilios Papadopoulos over at Georgetown, who's validating this assay in his lab.

And this is a dye. It's cationic. And it gets pulled into the mitochondria when the mitochondria have a functional electrochemical gradient, okay. What we're seeing here, which I'm pointing to, you're seeing mitochondria. These are Leydig cells, which are frequently used for steroidogenesis assays.

You're seeing punctate foci. That is a well-functioning cell. This is where we treated them with benzalkonium chloride, a surfactant, and it's got membrane activity. And it had a micromolar, where we're seeing swelling and bursting of mitochondria.

And this is what it looks like when there's an effect on mitochondria electrochemical gradient with JC-1. So, this is a candidate assay, and there's been many improvements on JC-1. There's JC-10. There's other ones out there.

Now I want to just jump into a case study. Go through this pretty quickly. This is with Anthralin. This is one of the 2,000 compounds that was screened through the steroidogenesis assay, okay. What you're looking at there is the MTT results. This is from Supplementary 2. And we're seeing no impact on mitochondrial function, up to a hundred micromolar,

even though this compound is known to accumulate and interfere with mitochondrial energy production. We're not picking up an effect at a hundred micromolar. That's a whopping dose. The pharmaceutical industry, when they screen, they never go above 30 micromolar because of cytotoxicity issues. So, this is a high dose.

What you're seeing here on the right is the analysis of all 11 hormones. This is the Mahalanobis distance. This is from Supplementary 7. We're seeing a response at 33 and 100 micromolar. That's a result of an impact on eight of the eleven hormones.

Anthralin appears to be a clear positive for disruption of steroidogenesis by looking at this. However, my question is, is this a direct or an indirect effect on steroidogenesis, this response we're seeing at the two highest doses.

We were able to find some JC-1 data that was done on Anthralin. This was published in a paper from 2012 in FASEB Journal. And what we're showing here, again, is this is the red. This is accumulation in the mitochondria. We're seeing a dose response. And in about one to five micromolar, we're

seeing disruption of the electrochemical gradient. The nice thing about JC-1, is if it doesn't get into the mitochondria, it gets cleaved from an aggregate into a monomer and you see green fluorescent. You can do somewhat of a mass balance. You can see as the red goes down, the green goes up. This is at a level far below where we're seeing an impact on steroidogenesis in this assay.

I wanted to mention that a lot of articles have come out on the ToxCast program. They've been very efficient and very prolific in publishing their work. There's dozens, dozens of articles out there. And one of the recent ones from Imran Shah looked at HepG2 cells, okay. They used high-content imaging to evaluate cellular phenotypic changes and to assess cellular state.

The purpose of this was to see can we equate cytotoxicity in vitro in HepG2 cells with a point of departure in in vivo study. It was a scoping exercise there. But they included the MitoTracker as one of the biomarkers that they evaluated in that study, okay. And MitoTracker, like I said, is very similar to JC-1. And they used this combined data to come up with a critical concentration, that tipping

point that I talked about. And I'll show that in a later slide.

But this is just pulling out the MitoTracker paper from the Shah versus Anthralin's steroidogenesis results. And what we're seeing here is an impact on mitochondrial function concomitant with an effect on steroidogenesis. So, the sensitivity is close. Granted, this is HepG2 cells. They're not H295R cells. But typically, you see some close cytotoxicity.

The JC-1 that I showed previously, that was with keratinocytes after one hour. And, again, I think that could be looked at as fairly representative.

This is pulling some of the data from Supplementary 7, where there's the Mahalanobis distances for the 600 odd that had dose response studies. We're looking here at chlorpropham, and what we're seeing here, this tipping point, this critical concentration, is below or just below where we're seeing a response on steroidogenesis, okay. That's interesting. That's telling me that this could be a non-specific effect. Chlorpropham is a herbicide.

In the Judson paper that was spoken to yesterday, that looked at the cytotoxicity measures, the 35-odd cytotoxicity measures to come up with Z scores, the herbicides were the good actors in terms of the classes of chemistry, when you look at the data. And that was one of the conclusions of the paper. It's because they target systems that are conserved in plants and generally not in animals. And that's why they were good actors.

Here's Volinanserin. This is a serotonin inhibitor. This critical concentration comes in right before the dose response. Here's another one, Isazofos. This is an OP, so it's an acetylcholinesterase inhibitor. Again, it comes in right below.

This is Propylparaben. This is a compound with a long history of safe use. This is in our shampoos. This is a preservative in foods. But it can be cytotoxic at high levels. Ten micromolar is a relatively high concentration, so we're seeing this effect.

And I didn't cherrypick this data. These are just four examples; you can go back and find many more. And the reason why you can do this with

the ToxCast Data is because these groups pulled from the same chemical library. So, there's a big overlap between what was looked at in the Shah paper and what was looked at in the steroidogenesis assay.

Here are a few recommendations. I think the program should compare MTT with assays that specifically assess mitochondrial membrane potential, before the assays used for prioritization. And, again, this is based on my experience in some of the data that I've shown you. And I think there's an opportunity to do this when the assay is reinstalled.

We heard the work was done at CeeTox. CeeTox is out of business. There's nobody running this assay right now. It has to be reinstalled somewhere, revalidated somewhere. So, there's an opportunity to go back and do this. And the white paper, again, recommends a quantitative method to assess mitochondrial toxicity.

And perhaps the cytotoxicity thresholds can be lowered with a better assay. The nice thing about MitoTracker and JC-1, and the likes of that, is you can see it. You can look at it quantitatively. You can look at it qualitatively. It's a nice representation. And the high-image capacity now will

allow you to take pictures of those wells and actually see what's going on.

Another important point, and this was discussed yesterday a couple of times, and that's characterization of the steroid hormone levels and kinetics should be done to inform a pathway analysis, okay. That's really important if we're going to use this to inform an AOP, we have to understand that.

And this was a comment that was made at the very first EDMVAC meeting when EPA brought this to the panel. Bill Kelse (phonetic) made this comment -- who's done a lot of work in this area. This is an adrenal cell line. It's being tricked into doing the steroidogenesis pathway, all the way through estrogen and testosterone with a unique media; and it's being induced with Forskolin to achieve the analysis of all those hormones for up or down.

And I think that that's necessary, but sometimes that approach can produce artifacts as well. I think we really need to go back and look at that. We really need to do that characterization before this assay is done.

My other recommendation is to initially only evaluate E and T for the EDSP purposes. And if

there is an effect on E and T, then you could perform pathway analysis. Right now we don't have an in vivo validated assay to go back and look at glucocorticoid responses. Yes, there's assays out there, but they have not been validated; things like the ACGH challenge test and other assays like that. And those are very difficult assays to do, if you've ever been involved with those.

I'm going to close with this final slide. And this is on the Statistics Charge Question 4. Mahalanobis distance is widely used in cluster analysis and various statistical classification techniques. It's frequently used to detect outliers that violate multi-variant normality. That's how you typically see it in the literature.

And I agree with the comment yesterday, that we probably should go back and check for normality. I think that's an important step. I think it can be used as an efficient approach, but maybe not as the only approach. A lot of people who teach this and use this approach don't use only Mahalanobis distance, they use another technique along with this, and there's other very similar techniques. And it

might be worthwhile just to go back and look at one of those techniques.

And as the last comment, EPA needs to demonstrate that the nominal Type 1 error rate of 0.01 should be confirmed so that the false positive rate is really 0.01. And I think 0.01 is a fair level. That's a typical alpha level used for outlier test normality test, not atypical.

I'll close there and pass it on to Brandy.

DR. BRANDY RIFFLE: Okay. I'm Brandy Riffle. I'm a regulatory toxicologist at BASF. And I will be presenting the comments on the Charge Questions 5 and 6 for the thyroid framework, and on behalf of the Endocrine Policy Forum.

I'd like to start off by repeating our commendation for the Agency and moving forward with this suite of in vitro assays; to start to prioritize chemicals for further additional testing for their inability to act with the HPT axis.

And to that end, I'd like to state that I think we can all recognize that the framework presented in the white paper is a bit of a work in progress at this point. Thus, I'd like for you to

take my comments, and our comments from the EPF, as suggestions to both the Agency; and to you as the panel, to work together, maybe to refine this framework so that it can become a bit more fit for its purpose, which is prioritization for the need for additional screening for interactions.

As we discussed with Charge Question Number 5, it's the completeness of the MIEs, the KEs, and adverse outcomes within the thyroid AOP. And Table 4-1 gave 15 molecular initiating events for thyroid perturbations. What we would like a little bit more discussion on is the relevance to each of these to toxicant effects in both humans and wildlife.

And though this is a low priority target for the Agency, pendrin, while it has effects in human pathophysiology, it actually has no known toxicant effects. So, maybe if we could discuss a bit more the relevance of each of these. And given these multiple MIEs, we'd like to understand a bit more how the Agency will begin to prioritize chemicals for screening and testing. Will they be using different potencies across the different MIEs in order to thoroughly capture everything?

Moving on again with Question 5. There has been some discussion on increased hepatic enzyme induction with the clearance of T4 and thyroid hormones. And we'd like to point out that increased hepatic enzyme induction is not an endocrine-specific MIE. It's a not specifically adverse, and it's a common finding in rodent toxicity studies, especially those done with high dose levels using the MTD approach.

Rodents also appear to be more sensitive to this MIE than humans, due to their different thyroid capacities. And interestingly, enzyme induction can be seen with a number of compounds that have no indication of specific thyroid or endocrine disrupting potential, such as the dietary constituents of cruciferous vegetables and one of my favorites, coffee.

Given these realistic exposures to chemicals, we'd like to point out that it's unlikely that toxicants would produce enough T4 clearance to produce deleterious effects in an animal. And, therefore, we question how the Agency will verify the relevance of this particular MIE for thyroid disruption.

Moving on with Question 5 where we discuss were there any missing pathways, adverse outcomes, or other AOP-related information.

As the panel and the Agency is aware of, the thyroid is an active area of research. And while T4 is a convenient measurement for researchers, the critical parameter is actual tissue levels of T3. Therefore, we'd like to suggest that the Conceptual Framework discuss factors that regulate tissue-specific thyroid hormone levels, particularly in fetuses, since this is a critical area. And these include increasing thyroid hormone synthesis release, blood transport protein, changes in tissue transmembrane transporters, altered intra-tissue deiodinase levels, and again, the metabolism or elimination of thyroid hormones.

Thus far in the white paper, we've had a limited amount of discussion on species-specific differences. And like Dr. Bever pointed out yesterday, there is an evolutionary conservation of T3 and T4 across species. However, there are differences, within these specifies, in actually the feedback mechanisms for the HPT axis.

A specific example of this is Corticotropin releasing factor, rather than TRF stimulates the thyroid pathway in amphibians. And these sort of species differences need to be accounted for. Additionally, even between mammals, we have differences in thyroid homoeostasis, such as the differences between rats and humans and their ability to handle excess thyroid hormone.

These sort of species-specific differences really do need to be accommodated in this framework in order for it to be successful.

With Charge Question Number 6, the panel was asked to discuss the importance of MIEs and biological and environmental relevance.

Table 4.2 organizes the endpoint data into two columns. You either have thyroid-specific endpoints or thyroid-related endpoints. And it's critical for us to point out here, and for the panel to recognize, that some of the thyroid-specific endpoints are subject to other stressors and generalized stress. And they include the decrease in thyroid hormones levels in rats, and the decrease in the developmental stage of the AMA. These two may not

be entirely diagnostic of anti-thyroidal activity, but maybe an indication of another confounder.

Other endpoints, such as asynchronous development in the AMA, were not included in the table. And the comparative thyroid assay was omitted. Therefore, we'd like to propose that Table 4.2 be revised to reflect this information, because the EPA white papers are used sometimes by other stakeholders. And it's important for the information to be clear and scientifically accurate in the papers.

As the panel pointed out yesterday several times, I think we had some great discussion, but we additionally wanted more information on the rationale for the ranking of the assays. But I think that discussion was held yesterday.

Moving on, we have the cell-free transport protein-based assays. And here we have a relatively high percentage of hits of the tested materials. We had 55 percent positive for interaction with TTR, and 40-percent positive interaction for thyroid binding globulin. And the question that we'd like for the Agency to consider, is whether they believe that greater than 50 percent of the chemicals that they tested had a meaningful impact on the

transport of thyroid hormone, that would produce physiologically relevant changes. If not, we want to understand how the Agency are going to use this data and interpret them correctly.

Again, as some of the panel members and the Agency knows, the thyroid, with chemicals that interact with it and affect it, it's sometimes difficult to identify specific MIEs in the pathway. Many of the agents, that alter the HPT axis, produce similar science in vivo, together, such as decrease in T4 and T3 levels, and the subsequent increase in TSH; changing in thyroid weights and the cell populations within the thyroid.

Therefore, we want to better understand how the Agency is going to validate these assays, given the limited knowledge of specific thyroid MIEs for many of these chemicals.

The simple overall conclusion is that the thyroid, and chemicals that interact with the thyroid, is a very complicated thing. And, therefore, the framework is going to need to be complicated and well thought out in order to be successful. We'd like to ask that once the HTP assays and a thyroid model is developed, that the Agency come up with an AOP-based

prioritization framework and a decision tree that could be fully evaluated by yet another SAP.

On the bright side, we, again, want to voice our support for EPA's pivot from the EDSP Tier 1 testing battery to ToxCast Tox 21 high-throughput screening methods; that way that we have a more efficient priority setting for chemicals that need additional testing. We really commend EPA's use of a systems-based model that takes into account many MIE's, life stages, different species, additional information, and critical parameters like toxicokinetics and actual exposure information.

We also support EPA's recognition that a single, positive high-throughput assay does not lead to a Tier 1 in vivo test orders. And the EPA, again I will repeat, is to be commended for their transparency. They have ensured this entire time that the data is publicly available. We can see their models, their codes, and everyone will have a chance to work with it and really go through it.

And we'd like to thank them for that and thank you guys for your time today.

DR. JAMES MCMANAMAN: Thank you. Are there questions for these presenters? Okay. Thank you very much. I have Dr. Esther Haugabrooks.

DR. ESTHER HAUGABROOKS: Good morning. My name is Esther Haugabrooks, and I would like to start my comments.

Good morning, Dr. Peterson, and the FIFRA SAP and colleagues. The Physicians Committee for Responsible Medicine is a national, nonprofit organization of over 150,000 doctors and laypersons, advocating for preventative medicine, good nutrition and ethical standards in medical research and toxicology testing. Thank you for the opportunity to comment here this morning.

According to the EPA's website, which we also discussed a little bit about this morning, the goal of the endocrine disrupter screening program is to screen chemicals rapidly for bioactivity and several endocrine pathways, while reducing the use of animals. It's commendable that the Agency has identified this need and provided resources to move away from in vivo testing, to smarter, better and quicker assessments, which is the purpose of why we are here today.

However, the rate of progress needs rapid improvement to realize this goal, while capitalizing on ever-changing science and technology. While progress since the EDSP 21 pivot has been promising, in the last two years, we have seen a disturbing lack of significant progress in implementing additional new tools developed by ORD at the regulatory level.

It is unclear how OSCP is facilitating necessary and important connections between ORD and OPPT, related to EPA's responsibilities under the EDSP. As it evaluates new tools, we urge the Agency to consider the validation status of the current Tier 1 in vivo assays.

When the Tier 1 battery was created, validation data for the pubertal assays were limited; and the Hershberger, these assays performed poorly. We asked, please do not set the bar higher, for more advanced mechanistic pathway models based on human cells, then it was for the assays that these models are replacing.

Under the EDSP, the Agency has always taken an AOP-grounded approach, and we encourage the Agency to continue in this direction. However, it is

not necessary to include every single molecular-initiating event, and key event, from all possible pathways in order to develop a protective endocrine screening system.

We need to learn from the information that has already been collected in this program, and focus efforts on what is needed to make a regulatory decision. Concerning the Agency's proposed high-throughput computational model for the androgen receptor pathways, we support the proposal, but are extremely troubled that the Agency is only proposing to replace a Tier 1 in vitro assay with this model.

It is perplexing that despite a robust AR model, when compared to a very limited validation of the Hershberger, that more progress has not been made towards replacing the Hershberger. We commend the Agency for making a clear statement back in 2014, that the 18 ER high-throughput assays, and an ER pathway model, would be accepted as an alternative to the three Tier 1 assays, which included the uterotrophic in vivo assay.

Hopefully, as a result of this current SAP, we look forward to the adoption and use of the AR model in place of the Hershberger assay. We ask the

Agency to devise a public plan to add another layer of transparency, outlining a roadmap to the placement of the Hershberger assay in the next fiscal year, along with its progress to develop an anti-androgen database similar to the work that was done with the uterotrophic database.

We encourage the Agency to participate extensively and, where they already are, continue participating in international forums such as the OECD. For example, the Agency's successful work on the ER model could have been proposed as a test guideline, harmonizing requirements to the benefit of other organizations which submit information to the Agency as well as other regulatory bodies. The Agency's work with the AR model can also be shared as a case study with international bodies.

As the Agency learns how to interpret and use data from new methods and approaches, these learning can be passed along to other countries so that international harmonization is achieved quicker. Likewise, concerning progress of the development of high-throughput model for steroidogenesis, we support the use of these models and hope that they will be

implemented to accelerate the pace of screening, while reducing animal testing and screening costs.

Furthermore, we appreciate the Agency's effort to compare the high-throughput H295R assay with the OECD test guideline 456. Yet again, we stress the need for the Agency to undertake the challenge of replacing in vivo test with new in vitro and computational approaches, rather than in vitro with more in vitro. And when this is done, to share their work in international forums.

This is another chance where we can ask the Agency to develop and publicly share a timeline for the development of models to replace other Tier 1 in vivo assays. Dr. Schappelle talked about it's going to take time; and so, we would like to see what that kind of time looks like. These Tier 1 in vivo assays would be such as the pubertal assays and the fish short-term reproductive assays.

Lastly, proposing an adverse outcome pathway-based framework concerning charge questions 5 and 6. For screening chemicals, it's a strong step towards 21st century science. Illuminating adverse outcome pathways for potential thyroid disruption is not only a way to move towards predictive science, but

also a way to harmonize methods and terminologies across international bodies. We encourage the continued development of MIEs and other KEs in language that could potentially be added to the AOP wiki.

It is critical that while developing a thyroid framework, the Agency avoid the compulsive need to be comprehensive. Gathering much data as possible has its place within research and is interesting to other stakeholders; however, we charge the Agency to continue to focus only on the information that is necessary and sufficient to inform regulatory decisions on thyroid pathways interactions.

In general, we have supported the EDSP 21, and are steadfast in doing so; but are just a little concerned with the apparent lack of progress of regulatory impact that OSEP seems to have made since 2014. It is important that continued development of faster, quicker and more reliable assessment is a mainstay within the EDSP program. And that those developments translate into changes in the Tier 1 and Tier 2 assays.

Therefore, it is critical that the EDSP remains a flexible framework to adapt emerging

technologies and key advances in science, without becoming stuck in endless validation exercises. We encourage continual engagement with interested stakeholders, in addition to devising a strategic plan that will track progress and document next steps for immediate replacement of the Hershberger and other Tier 1 in vivo assays.

Thank you for your time and thank you for listening. And thank you for consideration of these comments.

DR. JAMES MCMANAMAN: Thank you. Questions for this presenter? Marion?

DR. MARION EHRICH: Okay. You seem to be really concerned about too much validation? You have to know if something works or doesn't work.

DR. ESTHER HAUGABROOKS: I don't think the concern is necessarily on validation, as validation being bad. But just the standards, that we're placing against alternative methods, not be higher than what we've done for in vivo methods.

DR. MARION EHRICH: Okay.

DR. JAMES MCMANAMAN: All right. Other questions? Rebecca?

DR. REBECCA CLEWELL: I never know what's appropriate to ask. That brings up a good point. Maybe during discussion, can the Agency respond to these things, or no?

DR. JAMES MCMANAMAN: If it's related to the charge question. This doesn't sound like this particular question was.

DR. REBECCA CLEWELL: Okay. I will hold my question.

DR. JAMES MCMANAMAN: All right. Other questions? Okay. Thank you very much.

At this point, I think it's time to take a break. Is Catherine Willett in the room? She was scheduled to present.

DR. CATHERINE WILLETT: Yes.

DR. JAMES MCMANAMAN: You are here. Okay. Let's do a break first. We'll be back in fifteen minutes.

[BREAK]

DR. JAMES MCMANAMAN: Dr. Willett, if you are in the room. Thank you for accommodating the break.

DR. CATHERINE WILLETT: Thank you very much for the opportunity to comment. My name is Catherine Willett or Kate Willett and I'm here representing the Humane Society of the United States. I guess you could consider me also an EDSP groupie since I've doing this now for over ten years. And if I start to ask for signatures or autographs or something, then put me in a closet or something.

Yes, we very much appreciate this opportunity to comment. We created some written comments which were passed around this morning. What I'm going to do is just read some highlights of these comments. I'm not going to read every gritty detail. There's only a couple of things that are probably new conceptually from what we've already heard in various discussions and comments.

Some of the things that I have questions about in here were discussed this morning in the early discussion. That was interesting and clarifying. Some of this has already been addressed.

The HSUS, which is the nation's largest animal protection and scientific advocacy organization, we commend and support EPA's continued commitment to reduce and replace animal testing, while

improving the EPA's capacity to assess chemicals for potential endocrine activity. We are, however, disappointed in the slow progress that would replace the Hershberger. And I'll talk a little bit more about that in a second.

It's also concerning that much of the discussion suggests that it's necessary to develop or include in vitro assays for each and every of the animal endpoints, or key events in a pathway related to that endpoint. The critical question really is whether any potentially active chemicals would be missed, given the currently or foreseeably available spectrum of assays that are currently in the Tier 1 battery.

EPA's general approach to developing predictive models based on the ToxCast and Tox21 assays, include both productive and some potentially problematic attributes. There are a number of very productive attributes which include characterization of reference chemicals, which are critical for characterizing the assays as well as the model performance.

The curation of existing animal data to document the historical sensitivity specificity and

variability of that test. And comparison of the model with existing data for both in vitro and in vivo as was done with the ER predictive pathway.

The inclusion of caution flags to identify potentially problematic run data, including cytotoxicity. And allowing for inclusion of potential exposure information or modeling as part of a prioritization process.

Potentially problematic attributes include the choice of assays by availability rather than by design. And I'll talk a little bit more about that in a minute. The lack of evaluation of individual assays for performance and relevance. The lack of optimization of the overall prediction model to include only those assays necessary to maximize performance of the model. That was touched upon a bit by the first group of commenters.

Just as a reminder, the point of a screening battery is to flag chemicals of potential concern and not necessarily to characterize that concern. Acknowledging that screens should be designed to gain as much characterization information as practical and avoid excessive false-positives. In that vein, for future assay prioritization, it would

be most efficacious to identify and develop assays informed by critical key events that can cover a broad spectrum of biology in a limited number of assays. I'll mention that again in a second.

EPA is to be commended for exploring some of the potential strengths of the high-throughput assay format by including assessment of multiple steroid hormones in the HT steroidogenesis 295R assay. While the data analysis is at early stages, this type of thinking is needed to develop assays that address multiple AOPs involved in complex biological outcomes.

Regarding the thyroid pathway, EPA is also commended in requesting for expert input into the completeness of the current understanding of thyroid related pathways. To improve the predictive capacity of HT approaches, it's important to capture a full range of the relevant biology. Expert input is also critical for ensuring buy-in into the process of AOP-supported assay development and assessment.

EPA would also benefit from involving international experts in this conversation, who are engaged in similar processes, through the OECD or the European Commission. There are more specific comments

with respect to each charge questions. And I'll just briefly highlight some of those.

In terms of the first charge question regarding whether the HT AR pathway model could replace the single AR binding assay. Briefly, EPA has addressed the 2014 SAP suggestions by expanding the characterized list of reference chemicals, and by refining the predictive model. And has presented supportive evidence that the model predictions could be used in place of the current Tier 1 AR binding assay. While in addition, providing information about potential agonist versus antagonist activity. However, a few questions and concerns remain about the usefulness of this model.

Overall, the AR model performs as well or better than the AR binding assay alone. And unlike the AR binding assay alone, a benefit of the predictive model is that false-positives, due to cytotoxicity or nonspecific interference, can largely be identified via comparison with the ToxCast Tox21 cytotoxicity or proliferation assays statistical comparison between the model assays and a confirmatory assay to flag false-positive agonist activity.

Even though the AR model compares relatively favorably to the manual AR binding assay, the model relies on a large number of assays and statistical analyses, and seems excessive in order to replace one in vitro method.

At the SAP meeting to review the ER model three years ago, and in several presentations since, EPA has indicated that there is an ongoing review of the performance of the Hershberger Assay. We were a bit disappointed to find that the whitepaper mentions that this review is still ongoing and not part of the material for this SAP.

It's disappointing to see that EPA's predicting that even more HTS in vitro assays will be required to replace the Hershberger, since the current selection does not cover 5-alpha reductase or measure effects on other enzymes critical to steroid hormone synthesis.

The current AR prediction model is part of a battery of other assays. Including in vitro steroidogenesis and aromatase assays that do not also directly address 5-alpha reductase, but do address a broad range of steroidogenesis activities and other in vivo assays.

The question is not whether the prediction model can adequately account for all of the steroid biology potentially addressed by the Hershberger assay, but whether chemicals affecting androgen activity are otherwise affecting steroidogenesis would be missed by the battery and absence of the Hershberger. It's also worth considering whether the Hershberger adds value considering the performance of the assay.

During the validation of the Hershberger assay, it was noted that there was a high variability in assay results, some of which was due to the subjective nature of the scoring process. And so, an atlas was created to assist with this. However, it would be interesting to see if since the validation the performance has improved. At the time of the validation, the coefficient of variation varied by endpoint between 25 and 40 percent.

A recent optimization of the ER model has identified four assays that would provide essentially the same predictive capacity as the full 18 assay model recently published online this year. We suggest that a similar optimization be performed for the AR model pathway.

In addition, the utility of the model for prioritizing chemicals would benefit from the application of the integrated bioactivity exposure ranking. EPA has mentioned this also as something that they are intending to do. And we look forward to future integration and application of these approaches.

With respect to the charge questions 2, 3 and 4, regarding the high-throughput version of the H25R assay, I will just again read some excerpts of these comments. First of all, the description actually of the high-throughput H295R assay in the whitepaper was a bit difficult to follow because a discussion of the comparison of ANT version alone to the version that has all 11 -- it was kind of interspersed. And it was kind of difficult to follow the threads. It was kind of hard to figure out in addition to the complex statistics which were above my paygrade.

The bottom line in my reading of it, is the significance of the concordance of the HT assay with ANT with a low-throughput assay was not out of line considering the variability of the original H295R assay. That's really the bottom line.

The evidence presented in the whitepaper and included in the references, support EPA's conclusion that the HT version of the assay would perform as well, if not better, than the low-throughput version. The high-throughput version probably has fewer equivocal calls and fewer false-positives.

Although we are not in a position to evaluate the appropriateness of the Mahalanobis distance statistical approach, the idea of integrating measurement of 11 steroid hormones that addresses multiple related pathways into a single assay, is certainly a progressive step in increasing the cover of biological complexity in the HT assay format. Not only can the magnitude of the effect on these pathways overall be used to identify priority chemicals, the concentration response information on individual steroids is likely to be quite useful in unraveling mechanisms of action and in building quantitative models of the interrelated hormone pathways.

We also support EPA's suggested follow-on projects of continuing to identify appropriate reference chemicals for the predictive AR model, as well as for this model. And also, to include exposure

estimates or predictions in a comprehensive integrated bioactivity exposure ranking system.

With request to charge questions 5 and 6, this is on the thyroid pathways and AOPs. An OECD expert group convened a similar analysis of thyroid pathways and available related assays, which were published in 2014. The report also analyzed assays with respect to biological relevance and readiness. And prioritized assays for regulatory uptake.

This whitepaper includes many of those assays, not surprisingly since EPA participates in OECD expert groups and was likely involved in the report. And the reports served as a reference for the whitepaper. It also aligns these assays with an AOP framework. That not all of the assays are included is not surprising since many are still in development stages and/or are not amendable to the HT format.

Nor is it necessary to only implement existing assays. Development on application of thyroid-related AOPs, offer an opportunity to identify or create assays for a purpose.

For example, for a first-pass screen, it would be good to have assays that query key events that cover multiple AOPs, so that the first tier

screening covers all known related AOPs. Second-pass screening could then more extensively examine particular AOPs or molecular-initiating events. We hope EPA takes this opportunity to develop assays that fit this paradigm, including assays along the lines of what they are attempting to do with the HT295R assay.

We offer a couple of additional recommendations for developing HT models for thyroid activity. In addition to expert advice from this AOP, EPA could also consult international experts involved in similar efforts through the OECD National Coordinators of the Test Guidelines program. Or through the egg mass, for those of you who are on the egg mass.

And also, the European Commission who also recently published an analysis of thyroid-available assays and information. Through this framework contract they've had two, I think, workshops, and members of EPA have participated in those workshops.

Secondly, it's not necessary to develop assays for every MIE or to cover even most KEs. But rather to identify and develop assays informed by

critical KEs that cover the broad spectrum of biology in a limited number of assays.

Thank you very much for this opportunity to comment. And I'll take questions if there are any.

DR. JAMES MCMANAMAN: Any questions for this presenter? Okay. Thank you very much. At this stage, I think we can begin the charge panel questions and then break for lunch a little later. It seems a little early to go to lunch. If we can have the Agency come back to the table and give us Charge Question 1.

DR. SEEMA SCHAPPELLE: Just by matter of protocol, do we read the charge questions as we go?

DR. JAMES MCMANAMAN: Yes. You do.

DR. SEEMA SCHAPPELLE: Okay.

DR. RONNIE JOE BEVER: I'll read Charge Question 1. Please comment on the Agency's efforts to address the suggestions of the previous SAP. Thus, confirming the suitability of the current high-throughput androgen receptor pathway model to be used as an alternative to the low-throughput Tier 1 androgen receptor binding assay.

DR. JAMES MCMANAMAN: Okay the panel members on this charge question are Dr. Perkins, Berrocal, Pennell, Pullen Fedinick, Sobrian and Weller. Dr. Perkins is lead.

DR. EDWARD PERKINS: Thank you. This Dr. Perkins, Army Corps of Engineers. I'll try to summarize the notes I have so far from the people on my question. And if I miss something or am incoherent on what you think is important, please --

DR. JAMES MCMANAMAN: Ed, can you move the microphone a little closer.

DR. EDWARD PERKINS: Or if I don't speak close enough to the mic. In general, people appreciated the second-generation AR pathway computational network model. I think people agreed it was a nice way to integrate output from the multiple high-throughput assays. And they liked the efforts to develop the confirmatory in vitro antagonist assay data, and efforts for cytotoxicity information that you used to distinguish true AR pathway activity from biological and/or technology-specific assay interference.

This new model does seem to address concerns raised by the previous SAP for improving

scientific basis of the pathway model. However, while use of this model prioritize chemicals for testing is reasonable, there are remaining issues that should be addressed before it is used as the alternative for their LT Tier 1 AR binding assay.

Several people felt that you put a great effort together to address the comments. Particularly, with respect trying to look out for uncertainties, cytotoxicity and expansion of the assay battery and essentially the method to larger number of referenced chemicals.

One of the members felt that you did an adequate job responding to the concerns of cytotoxicity assay interference and transparency. And thinks that the use of the confidence score really helps this model be a major improvement over the last iteration of the AR model. And the additional assays probing antagonist behavior with some limitations -- which we'll mention later -- also appears to be a useful addition.

However, it was felt that the current AR model does not adequately address some comments on expanded chemical universe in some of the AR battery and non-classical AR binding issues. One member had

thoughts that this really does limit and prohibit the AR pathway model from being endorsed as an alternative to the LT Tier 1 AR binding assay at this time. And we'll expand upon that in a little bit.

On the limitations, I think we'll talk about those in each of the different areas. I've broken them down into kind of the principle areas that the SAP brought up. One was evaluating cytotoxicity. The issues related to cytotoxicity and cell stress were particularly important with respect to chemicals identified as antagonist. The model must be able to differentiate between cytotoxicity and cell stress and true antagonism.

Particular attention should be given to issues related to assay interference and to the factors in chemicals that contribute to cytotoxicity and stress. In general, members thought the Agency had done well in developing a caution flag or a cytotoxicity filter using cell-stress flags and other markers that were added to the model.

One question I had, as was brought up by one of the public presenters, was the issue of the mitochondrial membrane potential and on lower level chemical impacts on mitochondrial function, and how

that might affect dynamics of some of the cytotoxicity. That might be a consideration not just for the steroidogenesis, but perhaps for some of the other in vitro assays.

The Z-score approach, to flag the AC values considerably below medium AC 50, seemed to be somewhat informal but it does effectively compare the toxicity identified in the assays to the expected cytotoxic effects.

One member felt that although the results of the AR pathway model on the reference chemical list was quite impressive, comparison with results obtained by the Tier 1 binding assay indicate quite a disagreement between the Tier 1 binding assay and the proposed model. The Agency has given reasonable reasons for the discordance in the results. And while the justification is reasonable, it raises doubt whether this is a result of overfitting.

The AR pathway model has in some sense been trained using the reference chemicals in mind. And thus, the impressive performance of the model on the reference chemicals should be considered as some sort of in-sample validation or lack of independent test samples.

Although the Agency has tried to incorporate cytotoxicity and cell stress in the proposed framework, while also accounting for the additional source of uncertainty that cytotoxicity and cell stress introduction to the assay data, in one member's opinion the approach undertook for confidence scoring is not quite optimal yet and still required some work.

In particular, Figure 2-9 in the whitepaper was rather confusing as it showed a large spread of AUC values within each confidence score class. Ideally, it would have been better to have a greater separation between the difference confidence score classes.

Careful assessment of the general properties of solvent and test chemical in in vitro assay should be considered. This addresses one of the previous SAP questions. These factors are critical for AR bioactivity assays due to the predominance of chemicals that express antagonist activity rather than agonist activity.

Tier 1 AR binding assays do allow for testing chemicals that are water soluble. However, during the presentations, the Agency informed the SAP

that testing on water-soluble chemicals in HT assays has begun, will continue, but will be at a low priority. And we eagerly await to see what kind of results you get with water-soluble chemical testing versus using a DMSO.

Optimizing the assessment of activities -- and this goes back to what Kate Willett just mentioned on how many assays do you really need to assay before you get results, or how do you know what you're doing is sufficient. I think there needs to be a little bit more exploration of that. Do you really need all 11 assays to do this to get a similar answer there? Do different assays contribute more to the outcome than others? And this might help address some of the issues on interference. Do you really need assays that have significant interference?

The addition of confirmatory assays and orthogonal assays was a cleaver and effective way to address some of the issues of interference and having limitations of each individual test. The addition of two competitive binding assays seemed helpful for increasing the ability of the model to detect antagonist. But the assay still suffers from significant limitations.

In particular, the inability of the assays, to probe chemicals that are not soluble in DMSO, represents a significant barrier to the use of this model for screening chemicals that reside outside of ToxCast current domain of applicability. The addition of assays and/or results for chemicals in water or ethanol will help build confidence in the AR model's ability to replace or serve as an alternative to the current Tier 1 assay.

Expansion of the reference chemical AUC value range. There were questions, or suggestions from the last AOP, that EPA expand the range of chemical structure tested in the assay battery to maximize the screening potential, or understand the full potential of models that they're building.

One member felt that this was really adequately addressed by analyzing 1855 different chemicals of varying potency classes. Other reviewers felt that while the systematic review process for identifying chemical standards seem to be a robust process, and reference chemicals identified had a range of potencies, the current technical limitation context of the ToxCast system only work with DMSO-soluble chemicals, and make it impossible to determine

whether or not an AR model is suitable replacement for these LT rat prostate cytosol assays.

The current androgen receptor binding assay allows for use of ethanol water or DMSO solvents for chemical solubility. The chemical of the universe available for testing in the LT method is therefore necessarily larger than that of the AR model. Though the specificity, sensitivity and the BA -- balance accuracy -- are all quite high for the AR model. This calculation is only among chemicals that were tested in the ToxCast dataset, not in the entire EDSP universe.

Of the standards selected, there were only 31 in the 10,000 plus chemical EDSP universe. For example, Jarvis-Patrick Clustering, the Kmin of 5 and a K equal to 10, identified nearly 3,000 clusters across over 6,000 chemicals, including 6,425 chemicals in the EDSP universe with chemical availability. And 23 of the standards were not already included in the EDSP universe.

The selected standards, used by EPA, covered only 36 of the clusters identified. An examination of these clusters may identify further chemicals that could represent a broader portion of

its universe without trying to overwhelm you with numbers.

Solubility issues could make this model less able to identify chemicals with potential AR binding activity than the current Tier 1 test. Without the evidence of contrary -- i.e. that this model would be broadly applicable across the EDSP universe -- the inability to test chemicals that are amenable to the current Tier 1 test make this an unacceptable replacement, as felt by one member.

Demonstration of model reproducibility. The fact that the analysis incorporated several assays does support reproducibility of the results, in that it wasn't influenced by the sensitivity of one particular assay. Although, I don't know that this was really shown directly.

Using a bootstrap approach, you did show the interval estimates on the AUC, which reflect the uncertainty due to differences. You were able to generate interval estimates on the AUC, which were able to reflect the uncertainty due to difference across assays though.

While one member believed that the Agency has made a valiant effort in trying to

characterize the uncertainty AUC values derived using the AR pathway model, this member believed that more details are needed to understand whether the confidence intervals constructed using bootstrap resampling correctly account for all different types of uncertainties.

From the description of the bootstrap resampling procedure, it's unclear how the sampling is done, and whether the entire flow procedure, including model fitting to estimate the R values, curve fitting, et cetera, was applied. In particular, were the data relative to a chemical resampled within an assay and concentrations each time. Or was the data relative to a chemical resampled without doing the resampling within assay concentration pair.

There was a question on metabolic conversion of chemicals. Members felt that the bioactivity battery should include methods to assess potential effects of chemicals as well as the metabolites formed by enzymatic conversion in biological systems. I understand that EPA has plans of working with that.

In vitro assays may not always predict in vivo outcomes due to the limited coverage of

metabolic paralysis present in whole organism, especially important for compounds to undergo bioactivation; as these chemicals can produce false-negatives when tested in assays without metabolic activity.

This is a limitation of Tier 1 binding assays that should not be incorporated into the HT models as it's further developed beyond representing the Tier 1 binding assay. As I mentioned, the Agency recognizes the importance of metabolic-active cell lines and delineated is considering in silico approaches in additional assay with metabolic competency to address these issues in future planning.

Development of AR-related assays that do not follow classical genomic nuclear receptor pathways. The previous SAP asks the Agency to consider potential non-classical, non-genomic mechanisms that mimic or inhibit androgen bioactivity, such as non-DNA binding dependent pathways.

There are several ones including activation of second messenger pathways, including ERK, AKT, MAPK, that have been identified in a number of cell lines such as osteoblast and osteocytes that could be used. Indirect gene-trans repression could

also occur by the AR binding and sequestering transcription factor, such as activator protein 1, that are normally required to upregulate target-gene expression in the absence of AR binding to DNA. It was discussed that the Agency is considering non-classical and non-genomic mechanisms of AR pathway activation for future studies though.

While the Agency suggested it does not need to expand the chemical library to include non-genomic androgen antagonist, this decision seems to undermine the potential power of the tools they are creating and utilizing.

Since the goal of the EDSP program is to expand the use of AR tools to ultimately replace in vivo Hershberger assay, the ability of the model to identify chemicals that exert action outside of the canonical AR binding AOP is an essential one. And the Agency should continue efforts in trying to look at those.

Another challenge or request from the last SAP was to increase transparency in describing details about methods and results. Overall, we think the Agency has made a significant effort to really increase their transparency through publication of the

work and peer review literature, making data available to the public, and making protocols available.

All raw and processed data, as well as computer codes, are publicly available. And assays descriptions were described well in the supplemental files, including R code.

One thing on addition or optimization of the AR bioactivity test, as was mentioned before, many of these are based on availability. We would also encourage the Agency to consider more targeted development of assays, or picking assays based on key events in the chain of biological pathways that they're actually trying to look at, rather than just availability.

While one member understands that there is a point that there are maybe enough assays from a statistical standpoint, the Agency should made an argument that no key assays were made. Make more of a biological argument in setting up the AR pathway analysis that no essential events were made.

Replication. There was a thought that you should try to replicate known in vivo activity in in vitro assays. The EPA is currently developing assays to achieve this goal, thus the suggestion has

not been adequately addressed in this, but is in progress.

Mathematical issues. Compression of AUC scores. The AUC value range is narrow and lacks significant magnitude range for discriminating between AR bioactivity scores that are assigned to specific chemicals. The endocrine policy forum presents cogent arguments regarding this need to eliminate compression of AUC scores.

One statistical concern with the proposed approach, is the number of preprocessing steps involved in the analysis pipeline; which makes an inference procedure more prone to error and uncertainty, and may result in varying performance due solely to modeling decisions made throughout the pipeline.

Future iterations of the analysis approach may consider incorporation of other approaches such as the deep-learning approach offered by Borgen (phonetic) et al in 2017. It is noteworthy that development of this, and other approaches, is made possible by the EPA transparency, making assay data publicly available. EPA should continue to

strive for transparency in documenting and providing available data.

In regard to reproducibility, it was a little unclear to me on the performance-based issue if this really does apply to a model. I can see where the limited availability, or the high cost of 11 assays from several different companies for a small place, would be prohibited. But I would think that those companies had some validation test of their own, for those assays, that we could understand how this might impact further down the pipeline.

Additionally, I'm a bit unclear as to why the modeling that we're trying to validate -- the model predictions -- why this would not be accessible to a normal model, using other assay data, to see whether the pipeline gave similar results or not. I'm still a little unclear why performance-based validation is more appropriate than some of the traditional portions. Especially for modeling where you're holding out one set of data -- do the development of the model on one set, and then you test the set on a dataset that you haven't actually incorporated into the model or used to develop the model.

I think that's more or less everything that I had gotten from everyone and random thoughts too. We'll open it up to for anyone else to contribute.

DR. JAMES MCMANAMAN: Okay. Dr. Berrocal.

DR. VERONICA BERROCAL: I don't have anything to add. Ed has said everything I sent him.

DR. JAMES MCMANAMAN: Dr. Pennell.

DR. MICHAEL PENNELL: I would just like to add to the concern about the number of preprocessing steps in the analysis. Perhaps some of those issues could have been addressed in their bootstrapping. But it needs to be made clearer, exactly what they were doing in the bootstrapping.

DR. JAMES MCMANAMAN: Thank you. Dr. Pullen Fedinick.

DR. KRISTI PULLEN FEDINICK: The other issue that I thought would be helpful to think about is the noncompetitive mechanisms of antagonisms. That wasn't covered in the new assays that were added. And not having this could render the competitive binding assays less useful than proposed; and could significantly impact the ability of the model to

correctly identify chemicals that act in non-classical ways.

And so, in addition to looking at the chemicals that could fall within the non-genomic or classical antagonisms models, really thinking about the ways in which chemicals can act as antagonist as well. I'm thinking about additional assays that would cover that.

It would also be useful moving forward for the Agency to explore the use of higher maximum concentrations, in order to reduce the false-negative rate found during the comparison of the Tier 1, List 1 results to the AR model. If you don't have the technical limitations of going above 100 micromolar, it would be interesting just to see for chemicals, that you saw the false-negative for at least. And expending that potentially based upon chemical similarity to see if similar results would be found. Or just doing that for a subset of the chemicals within the ToxCast universe.

Also, the program should also ensure that it's being more sensitive than they are specific. The EDSTAC recommended, in their report in 1999, that the assays have the primary objective and the

minimization of false-negative or Type 2 errors, while permitting an as-yet, undetermined, but acceptable level of false-positive or Type 1 errors. So, ensuring that your Type 2 errors are very low.

And the Agency's response to this question yesterday about false-negatives being allowed due to this being a prioritization, is in some ways misleading. In that these tests will not only be used for prioritization, but also for screening. And so even though we've been talking a lot about prioritization, these are potential replacements or other ways in which to submit information for the Tier 1 screen. And so, ensuring again that those false-negative rates are very low is essential.

The fact that this charge question specifically asked about the ability to serve as an alternative to the Tier 1 screening test, makes this even more important. The inability to evaluate the chemicals that reside outside of the chemical standards tested, limit the confidence in this particular method.

And then finally for reproducibility, it seems as though -- and this is kind of brought up in some ways as well -- but asserting that orthogonal

assays demonstrate reproducibility in some ways seems a flaw in logic. Particularly, since each of the assays is used to create a composite score for each chemical. And so, identifying ways to assess reproducibility in naive labs may not be necessary as what was done with the OECD validation processes.

But the reproducibility of this model outside of the Agency, or with non-ToxCast data, hasn't yet been demonstrated. It would be interesting to explore ways in which reproducibility might also be addressed outside of the ways that it has been so far.

DR. JAMES MCMANAMAN: Dr. Sobrian.

DR. SONYA SOBRIAN: I have nothing to add because Dr. Perkins has adequately incorporated my comments.

DR. JAMES MCMANAMAN: Thank you. Dr. Weller.

DR. GRANT WELLER: My comments were captured by Dr. Perkins. I would just add that from the perspective of the data product involved here, so the analysis pipeline and the mathematical model, I do think the concerns of the previous SAP have been adequately addressed. And I just repeat other panelists remarks that the EPA's efforts towards

transparency and reproducibility have been really impressive.

DR. JAMES MCMANAMAN: Okay. This charge question is open to other panel members if they would like to make comments. Yes, Dr. Ehrich.

DR. MARION EHRICH: I have just a question for the people that answered this. There seems to be a lot of worry about water solubility. Yet by the time you put the chemicals in the assay, you can't have more than .1 percent DMSO so essentially, they do have quite a bit of water solubility. This came up again and again and I just wonder why.

DR. KRISTI PULLEN FEDINICK: I don't think it's necessarily water specifically, right. It's just using water as an example of the limitations of the current ToxCast system. There are issues with medals, there are issues with volatile chemicals. The ToxCast process in and of itself is limited in its ability to probe the full 10,000 chemical EDSP universe.

I wouldn't focus necessarily on water so much, that's just one particular media that would be important, I think, for a lot of these chemicals.

But ultimately, it's really about ensuring that we have coverage of all of the chemicals that are required to be tested under this program. And so, I think that's really more of the issue, rather than water specifically.

DR. MARION EHRICH: That was the word I heard.

DR. KRISTI PULLEN FEDINICK: Yeah. It just came up as an example, right. Just indication of the smallness of the universe. And so, when we look at that big graph that has little tiny bits that said -- you know, or the big circle and the little tiny circles that say, this is what we covered so far. You'd have a slightly bigger circle maybe that covers 10 percent of that larger EDSP circle that's currently within ToxCast testing ability.

And so, we're not looking at just 10 percent of the EDSP universe. We need to be able to explore 10,000 chemicals if not more. Again, water is just one potential limitation, but you could substitute that with volatility and other types of chemical characteristics.

DR. JAMES MCMANAMAN: Dr. Ehrich.

DR. MARION EHRICH: I would suggest that maybe that be put in the response, so it doesn't sound like water solubility is the primary concern here, it's something else.

DR. KRISTI PULLEN FEDINICK: Yes. That's great. It's in there too, so yeah; we can do that. Thank you.

DR. JAMES MCMANAMAN: Dr. Clewell.

DR. REBECCA CLEWELL: I'd like to respond to that specifically. And then I also made some notes. I wasn't on this team, but I'm obviously very interested in the topic.

I think it's important to point out that this domain of applicability for in vitro assays is broadly true for all in vitro systems. And including the low-throughput assays that we are specifically trying to replace in this charge question.

The question of whether this assay -- the low throughput can be replaced with the high throughput, that doesn't have anything to do with the domain of applicability for in vitro assays. The issue you bring up is important, broadly, for in vitro replacement for animal testing, but it doesn't

preclude this assay from being replaced with another in vitro assay.

They all have that same issue in that it's difficult to use volatile chemicals. It's difficult to use medals. It's difficult to use anything that's highly lipophilic. Anything that binds strongly to proteins.

There are a lot of kinetic issues in vitro and there are some very brilliant people working on that issue and the EPA has taken on – NCCT, in particular, has taken on several of those issues in terms of metabolism. And I understand they are actually looking into lung systems for volatility.

I think that that is an important issue, but I also think we need to not confound the question that we were given with that particular issue. Because it's just not a fair argument to use against one in vitro assay as a replacement for another.

I would love to talk about domain of applicability though. And I would love to hear, either in the response from the Agency or I'm not sure how this goes, a plan forward for how we can prioritize. And say for chemicals that are useful,

can be tested in an in vitro system, then this would be our plan forward.

For those chemicals that can't, we still have a plan forward. And maybe it's the old fashion way. And maybe it uses more animals than we would all prefer to use, but at least we have a path forward. And we're not going to just say we'll never test them because we don't know how to use medals in an in vitro system.

I'm sure the Agency is thinking that way because they're not going to just say well we're never going to test medals. I think there's opportunities to have the discussion; I think it's an important discussion to have, but I don't think it rules out the utility of the current in vitro assay as a replacement for another in vitro assay.

In terms of the non-genomic signaling, I too agree that it is very important. Pretty much every nuclear receptor pathway has non-genomic signaling that are important to the overall phenotypic response, which is overall what we would like to avoid. It's true for ER, AR, thyroid, CAR, PXR. Keep naming receptors, you'll find a non-genomic signaling pathway that's important. And it's not sufficient to

just say can a chemical bind a receptor and then, yes I can guarantee that there will be a phenotypic response.

On the other hand, again, our charge question is can we replace an AR binding model with an AR binding model? And I would say yes. As a matter of fact, not only did you say can I replace an AR binding model with the binding model, but now you're replacing it with a pathway model that at least moves us beyond the binding to the dimerization and the transactivation. And the transactivation in itself covers some of the concerns, I think. And it could be tested regarding non-competitive binding in particular.

There are ways that these could be tested. I agree that it would be useful to go back and evaluate whether the transactivation assay could account for non-competitive binding. I agree that I can't think of a non-competitive binder off the top of my head. But if people know of it, then I say we put it in a document and we ask to see how that performs in a transactivation assay.

The reason I bring up the non-genomic signaling -- because I do think it's important,

particularly if you want to move to replacing the Hershberger or any of the in vitro assays. I was very pleased to hear that there is a proliferation assay in development. Or even maybe in testing. That brings a phenotypic response into the suite, which is really important.

Because to get a full proliferation response in response to androgen receptor binding, you need a lot more events happening within the cell than just a binding event, or even just a DNA-binding event. You need to have a concerted cellular response to the androgen ligand. And so that's a very important addition to the suite. I would recommend that that be added into the model as soon as possible, possibly before making it useful for the EDSP.

I think that's most of it, except the optimization of the model to see whether we can't remove some of those. There is a lot of redundancy in the current model. There is more than one binding assay. There is more than one transactivation assay. Could we reduce that and provide performance-based standards for whatever transactivation assay you were going to use. And whatever binding assay you are

going to use. And make this actually more feasible for companies and smaller businesses to use.

As a small business, I both run assays and I contract out assays, and it is not feasible for me to contract out 11 assays, it's just not. But I am I'm confident -- though I haven't run the numbers with the AUC calculation myself -- I'm confident that there's probably performance-based standards that could be applied to the general concepts of the assays that are being used as part of the AUC, so, the binding or the transactivation; that once they're applied, it wouldn't really matter if I use your transactivation assay or my transactivation assay.

I can use a transactivation assay that is appropriately sensitive, and that should be enough to get me the data I need to do an AUC model. And I think that's just tremendously important. Because we don't all have the resources that the government has. I can't do 11 assays for every chemical I've got an interest in. But I can do a binding assay and a transactivation assay.

I'm almost done, I swear. But I actually think it's really important -- and the reason I wanted to speak up. I think it's important that we

take this opportunity to support the replacement of a low throughput, kind of hinky, AR binding assay with a comprehensive suite of assays. And then we provide recommendations for how it can be better.

Let's take a step forward. Let's provide recommendations for how it can better and let's improve as we go. Maybe as a panel we can provide some -- my interest was piqued with the comment earlier about why aren't we at least trying to replace the Hershberger. We've done it for the uterotrophic assay, right, so what is different about androgen receptor. I think having the proliferation assay will make it a much easier move to an in vivo replacement for at least one of the short-term in vivo male rat androgen receptor assay.

But I think it's important to highlight that this is actually a good replacement for the current in vitro, and I don't think we're that far off from the in vivo. At least the short term. And I'm going to leave it at that.

DR. JAMES MCMANAMAN: Thank you. I think Veronica had her hand up first. Dr. Berrocal.

DR. VERONICA BERROCAL: Yes. I guess I have a question. And I'm speaking as somebody who

doesn't understand assay very well. I guess I'm confused about what is the goal of this charge question and what is it that we're trying to do here. And whether this androgen receptor pathway model is something that the EPA should use for prioritizing chemicals or companies outside of EPA should use.

And the question is -- I think the reason why I have this question is because some comments that have been raised during the public comments about reducing the number of assays. When the SAP gave a suggestion to EPA in the previous meeting -- in the 2014 meeting -- to actually increase the number of assay. And I think that the suggestion was given because the idea was to have EPA be as thorough as possible.

I just feel conflicted in the sense that this question is asking whether the EPA has addressed the comments that the previous SAP has raised. And instead we are receiving comments about reducing the number of assays to achieve another goal, which I don't think is the goal that this question was trying to address. But maybe I'm misunderstanding the charge question.

DR. JAMES MCMANAMAN: Thank you.

DR. REBECCA CLEWELL: Can I respond to that? Just shortly.

DR. JAMES MCMANAMAN: Sure.

DR. REBECCA CLEWELL: I wonder if -- and I wasn't part of the previous SAP. But they did ask for more assays. Within that same kind of context, they had mentioned the non-genomic signaling. And I wonder if the request wasn't more about including assays that address different parts of the pathways. Because really, if you already have two binding assays, do you need five? More isn't always better. Did they address it? Sure, they have more assays, but that could be separate from optimizing the assays.

DR. JAMES MCMANAMAN: Dr. Perkins.

DR. EDWARD PERKINS: No. It really wasn't that. It's that there were feelings that endocrine receptor membrane signaling might have impacts, rather than translocating to the nucleus. That was more of the directive than needing more assays along the same pathway. It was kind of a different thing.

DR. REBECCA CLEWELL: See, that makes sense. And then that would mean we wouldn't want to

add more binding assays or more transactivation assays. You would want to add assays that specifically address that question.

DR. EDWARD PERKINS: But that was the ER pathway one, which was looking at a much larger scale. Right here, they've really focused it on trying to replace this binding assay, which I think is fairly appropriate. It's not going beyond knowing essentially have you activated the binding -- replacing what the prioritization with the binding assay is; it's, is it interacting enough with AR to activate downstream events potentially. I think it does do quite a bit of that.

DR. JAMES MCMANAMAN: That was Dr. Perkins and Dr. Clewell. Dr. Pullen Fedinick.

DR. KRISTI PULLEN FEDINICK: Just to go back to the comment about the demand of applicability and the inability for in vitro assays in general to measure things like medals, and I think that's absolutely true. But with the ToxCast assays, so a chemical that just comes to mind is glyphosate, for example. And EPA should certainly correct me if I'm wrong. But under the current ToxCast assays that we have, glyphosate wouldn't be able to be run with this

assay, but would be able to be run with a low-throughput assay.

And so, it's not a medal. It's not lipophilic. It's not all these things that you mentioned. I think that not having the ability to probe something that is largely present in the environment, something that is water soluble, something that potentially should go into a prioritization scheme. Right now, that would even fall outside of the prioritization context of these tools completely.

And so, you would have chemicals that would just be sitting in a holding bin until the Agency is able to develop something to address some of these issues. And so, wouldn't then even be prioritized to go into further testing. I think the problem of solubility, or the problem of the limited demand of applicability, isn't just those things that would fall outside of in vitro assays in general, but specific to ToxCast.

DR. REBECCA CLEWELL: I think maybe you and I heard two different things yesterday; glyphosate because it's not DMS soluble. What I heard yesterday is there's no reason that water soluble chemicals

couldn't be used with a water vehicle in the ToxCast assay. It just hasn't been done yet. But it's not limited and unable to do that.

And then the other thing I heard yesterday, that's helpful in this situation, is that this isn't necessarily a replacement, it's an alternative. If they're not able to test in the ToxCast assays, they could still use the old binding assay. There's no limitation to having this model on the table in addition to the old model.

DR. KRISTI PULLEN FEDINICK: Just to comment back on that. I think that the Agency having done the water test, I think is amazing. I would love to see those, but we've not seen them. And so, if what we're supposed to be analyzing is the underlying science, we can only analyze what it is that we've seen. Had the water test been presented in this meeting, I think I would be much more comfortable saying that this is applicable.

We could use this to prioritize chemicals that are then going to go on to further testing. Or to use within our scientifically relevant information. But since we haven't -- as a panel, we don't have access to that information currently, we

can't then evaluate the applicability beyond what we've already seen.

Again, because these tools are being used for priority setting, and for testing, I think that we have to think about the pools that can even then be brought into those Tier 1 tests in the first place. You want to make sure that that prescreen, that prioritization cast, is as wide a net as possible. And then you can then prioritize from that wide net that would then go through the EDSP screening process.

But if we have a very narrow net in the very beginning, that means we're screening increasingly smaller numbers of chemicals that are, again, missing chemicals that are relevant to human exposures.

DR. JAMES MCMANAMAN: Okay, thank you. Dr. Pennell.

DR. MICHAEL PENNELL: A comment on the request to reduce the number of assays. I kind of made this comment that Dr. Perkins summarized. I kind of like the idea even if you do have different assays that are measuring the same thing. One thing that does eliminate is having to undergo additional

uncertainty analysis later on, right. What you're presenting is dependent upon one particular assay that you chose for each of the binding sides, for instance. The fact that you're using multiple sources of information, and getting estimates, which are kind of averaged across, I feel is a strength.

DR. JAMES MCMANAMAN: Thank you. Dr. Perkins.

DR. EDWARD PERKINS: I just have to support that again. The use of orthogonal assays, I think, really helps a lot. It's much like this panel; we have to have statisticians to compensate for the biologists. I think it evens out in the end.

DR. REBECCA CLEWELL: That is more assays. It's having more assays within that pathway. And it actually gives you more information than having a lot of assays for just two key events. I'm not against having more assays or more information, but it's always a balance of what can we actually do, feasibly, within the resources that are available. Dr. Paul is probably liking that I'm saying that. And then also what you need.

What the EPA, I believe, is doing -- though I'm not privy to it -- is optimizing the

estrogen model in that way; in saying if we have these number of assays, how can we reproducibly get the same quality of results. So, the same balanced accuracy, or at least close enough, if we pull these pools.

That's a statistical exercise I bet you would probably be really good at; is saying how can we ensure that the smallest number of assays, with the greatest payoff, and that's just optimization. I'm not saying just knock out important assays, but I think it's important to say how many do we really need to be consistently able to predict a response.

DR. JAMES MCMANAMAN: Is that part of the charge question?

DR. REBECCA CLEWELL: Yes. Inherently.

DR. JAMES MCMANAMAN: I think it's an important discussion, but I don't know whether it actually -- I'm sorry. That was Dr. Clewell and this is Dr. Perkins.

DR. EDWARD PERKINS: That actually does kind of falls into -- reviewing the optimization of the assays and the pathway was one of the SAP suggestions. And this kind of falls into optimization. What is the optimal combination of

assays in this work to get desired outcomes? It kind of fits in there, I think.

DR. JAMES MCMANAMAN: Dr. Pullen Fedinick.

DR. KRISTI PULLEN FEDINICK: One of the questions I would have in terms of the charge question is, is there a scientifically justifiable reason for optimization, rather than a resource specific question. And I think that if what we're here to address is the science of it, then potentially having more -- if we, again, think about this in a completely resource rich environment.

And again, we're thinking about this in a vacuum in a way, just looking at the science, then the number of resources, or the amount of money that it cost to run a particular assay, isn't necessarily a scientific concern; but a financial concern that then would be addressed after this. The Agency then has to make a decision based upon those types of question, but that's not for a science advisory panel to discuss.

DR. JAMES MCMANAMAN: Okay. Other comments? If not, I'll send it back to the Agency and ask if everything was understood.

DR. SEEMA SCHAPPELLE: A couple of points of clarification and a question. And I'd also like to turn it over to Dr. Bever for clarification on the charge question. And Dr. Judson for some additional issues as well.

I do want to underscore the point that what we are posing is an alternative and not a replacement. And this is still proving the ability -- and this is a comment that was made earlier here -- providing the ability to use the Tier 1 assays for evaluation when it's appropriate. And I think that really touches on a lot of the issues that we've talked about today. Things like the domain of applicability and some of the prioritization. This does not preclude us from prioritization. It gives us an additional tool. That's one thing I wanted to bring up.

Also, a comment was made earlier with regard to analysis of our chemicals in terms of a low-throughput comparison to a high-throughput comparison. And just a reminder, with our List 1 chemicals with our test set of chemicals, they were evaluated both in low throughput and high throughput, providing for a dataset that's valuable and useful here.

Question about the Tier 1 AR-binding versus the high-throughput alternative that's being proposed. Recognizing there are limitations with the high-throughput approach. In many cases some of the examples that were cited, metabolic conversion and others as well, those are current limitations in the low-throughput analysis also. Just some things to keep in mind there. Maybe that one's not so much a question after all.

Let me turn it over to Dr. Bever for additional clarification on the charge question. And Dr. Judson, can I invite you up to address a couple of issues as well.

DR. RONNIE JOE BEVER: Okay. Well, we actually had a meeting to make sure the charge questions were clear. And to just reiterate, once again, what Dr. Schappelle has said, we're proposing this as an alternative. We're asking for comments and suggestions, but the real question is, can this serve as an alternative?

One of the high-throughput assays in the model is actually kind of like an upscale version of the low throughput. Also, Dr. Judson had mentioned that -- I mean, we're dealing with media. It's water.

Most of these water-soluble compounds, it should go just fine. We tested in DMSO, that carries a lot. And with cell systems -- and some of these assays are cell systems -- you can't necessarily use strong solvents. I mean, the other choice for the low throughput is ethanol. I don't really see that there's going to be any stream domain restriction here, but we are carrying out some of that investigation.

I'd like to point out, too, that the low throughput is not necessarily a gold standard. It is a validated study, we did comparisons with it, but that doesn't necessarily mean that the low throughput is superior to the high throughput. I really feel like it's very important to have reference chemicals come about through a systematic literature review.

And to talk about that domain of applicability, you realize that receptors require a certain structure to actually bind and elucidate the effect. Not every class of chemicals are going to be able to do that. It's just not going to work like that.

The androgen receptor and an estrogen receptor specifically are extremely important

therapeutically for various reasons, including hormone therapy of cancer. The point there is that there's been a lot of bright scientists looking into developing chemicals on these receptors. Thereby our systematic literature review was looking for any examples that we could use for our purposes as reference chemicals.

We feel like there has been no major losses like chemical classes at this stage. We feel like we have a very strong contingent of reference chemicals. At this stage, we're only proposing as an alternative for the low throughput. That doesn't mean that the Agency's efforts here are done.

As I described, there's a process here. We're going to further optimize; we're going to look into ways of integrating our assays, or adding more to our assays, so that there could be potential alternatives to, first of all, the Tier 1 in vivo. The Hershberger was mentioned. Of course, that's on the radar.

DR. JAMES MCMANAMAN: Dr. Bever, let me interrupt here. The point of this is to find out whether the panel's comments were clear to the Agency,

and not to have the Agency defend their approach. Because we could be here until this time next year --

DR. RONNIE JOE BEVER: Okay. I'm sorry.

DR. JAMES MCMANAMAN: -- if we did that with every question.

DR. RONNIE JOE BEVER: I was just thinking there was some misunderstanding, so what are we presenting. I'm sorry. I'm through. I'll pass it to Dr. Judson.

DR. JAMES MCMANAMAN: Same thing for Dr. Judson.

DR. SEEMA SCHAPPELLE: Dr. Judson, there are a few issues that I noted down. Can I maybe just make sure that those are on your radar, as well, as we get going.

DR. RICHARD JUDSON: Yes. Just so long as I don't break the rules.

DR. SEEMA SCHAPPELLE: That's right. Please correct if I'm breaking them as well. Redundancy in the assays was brought up, higher concentrations for false-negatives, and then some of the issues with solubility. Hopefully, those are on your list as well.

DR. RICHARD JUDSON: Well, I guess what I would say is all of the comments and questions from the committee were -- I disagree with many of them, but they were all very clear. Sorry.

DR. JAMES MCMANAMAN: Perfectly fine answer, thank you. Okay, with that then, I think it is lunchtime. And so, we'll come back to charge question number 2 at 1:05.

[LUNCH BREAK]

DR. JAMES MCMANAMAN: To return to the task at hand, we're on Charge Question 2. And I've been asked to remind the panel that we had an open meeting in which we went over the charge questions to see about their clarity. I hope that that wasn't for naught because there seems to be many questions about clarity. Let's see if we can get through the questions on the scientific merits.

Charge question 2, if we could have the Agency read that into the record please.

DR. RONNIE JOE BEVER: Charge question 2; based on the comparison of the performance of the high-throughput H295R assay, with the low-throughput

H295R assay and the effects of reference chemicals on the synthesis of testosterone and estradiol levels only, please comment on the suitability of the high-throughput H295R assay as an alternative to the low-throughput H295R assay. See sections 3.3 and 3.4.

DR. JAMES MCMANAMAN: Thank you. The panel for this is Dr. Belcher, Dr. Clewell, Jett, Nagel and Pullen Fedinick. Dr. Belcher is lead.

DR. SCOTT BELCHER: I'm going to go ahead and read a summary of the comments that were incorporated from the group members. And at the end, I'll ask Dr. Jett, who is on the phone, to add his comments that I have here on my phone, but were not incorporated into what I have. And these are summary comments. And the other panel members as well, please, if I get anything wrong or don't properly reflect the information that they've given to me.

The high-throughput H295R steroidogenesis for the measurement of E&T only, is felt to be based on generally well-conceived modifications of the existing validated low-throughput H295R cell-based steroidogenesis assay. This assay was modified to facilitate the analysis in 96-well cell culture format. Along with qualitative steroid

assessment and HPLC tandem-mass spec analysis of these multiple steroids.

Conceptually, the high-throughput H295R steroidogenesis assay is a scientifically sound alternative to the low-throughput assay. While there was a range of opinions expressed regarding the current ability of the high-throughput assay to substitute for the low-throughput assay, there were some important limitations identified that would not allow substitution of the high-throughput steroidogenesis assay for the low-throughput assay at this time.

The high-throughput H259R steroidogenesis assay -- which I'll probably for brevity start to just refer to as the high-throughput assay -- benefits from several strengths. Although there were also some specific concerns related primarily to sensitivity and reproducibility that were judged to limit the suitability of the high-throughput steroidogenesis assay as a replacement for the low-throughput assay for E&T.

It's felt that the incorporation of forskolin pretreatment to increase baseline steroid production in the assay was a positive modification

and a major strength for increasing the throughput. There was, however, some points raised regarding the findings for the use of forskolin as a test compound in the comparative analysis between the assays.

As it was done in Karmaus et al 2016, this would complicate this comparison. As was described, the high-throughput assay affects for forskolin must have been compared relative to the DMSO only baseline. Whereas, the other reference compounds would be compared to a forskolin pretreatment baseline.

Moving on, retaining an assessment of cell viability as part of the assay was also considered a strength. Although, the reduction in cell viability standard, from the 80 percent in the low-throughput assay to 70 percent, is considered worthy of further evaluation.

It is appreciated that 70 percent viability was indicated as the statistical limitation of the assay for this application. It is also considered likely that a 30 percent loss of viability could be biologically meaningful and likely to impact the assay results.

Additional justification for the appropriateness of the 70 percent viability cutoff is recommended. Along with the information supplied today, recommendations were put forth for evaluating the impacts of the findings on the results if the viability cutoff were to be set at 80 percent as it was in the low-throughput assay.

An additional suggestion was that possibly defining and investigating the utility of an appropriate cytotoxicity Z-score, as was done for the AR assays, should be investigated.

Additionally, the use of an alternative cell viability assay that is independent of mitochondrial reductase function was suggested. Another suggested characteristic of this alternative assay is that it might be less variable than the currently used MTT viability assay.

It was stated that there was a potential that uncoupling cell viability assessment from mitochondrial function may be more important, for assays evaluating steroidogenesis, than it would be for other endpoints. As was indicated earlier today, much of the steroid metabolism occurs in the mitochondria. It may also be valuable to examine and

consider viability and mitochondrial function independently.

It was also stated that for the set of reference chemicals used in the inter-lab analysis of the OECD guideline low-throughput steroidogenesis assay, the high-throughput assay appears to be performing with relatively less sensitivity. The sensitivities reported in Figure 3.8 of the whitepaper were found not to be acceptable for public health protection. And that the failure of the high-throughput H295R assay to accurate identify the estrogen and testosterone production disrupting reference chemicals was considered to render the assay, in its current form, inadequate to protect the health of populations.

There were some additional concerns related to replication or reliability in the high-throughput assay, and some of the approaches used for the comparative analysis. Firstly, it was not readily apparent if the performance of the high-throughput assay in an intra-laboratory performance assessment, across seven different laboratories, was the most appropriate matrix for the comparative evaluation. I

may have misspoken and said high-throughput, but I mean the low-throughput in that statement.

There was identified a general lack of replication, both technically and biologically. With the information presented, it is not possible to interpret the liability of these tests from run to run without additional specific information regarding the consistency of the results across replicates. The reliability of the assay analysis from day to day, across blocks, should be established. This concern extends to the ability to replicate results for future testing; as it was referred to earlier today, the concept of transportability.

It would have been useful, for example, for the Agency to report the independent retesting of chemicals, or a subset of chemicals, tested in the Karmaus paper to assess replicability across the time domain. Related more specifically to the comparative analysis of E&T, while it is indicated that 16 percent of the screened chemicals were analyzed in more than 1 plate block, it is not indicated how many times the individual reference chemicals were analyzed.

Most test chemicals analyzed in the high-throughput assay were examined only once as

duplicate technical replicants in a single block. But one is left to assume that this is likely not the case for each of the reference chemicals. This is because the reference chemicals were analyzed by ANOVA and Dunnett's test for comparison with the low-throughput assay results.

It would therefore, thus, seem reasonable to assume that more than one biological replicate was analyzed. But the lack of specific information on the biological replicants makes it difficult to really compare the reproducibility of the results of the reference chemicals.

A few more general comments. There was also expressed, some concerns related to the inability to fully assess the appropriateness of the prescreening approach that was used. It is stated that the Karmaus paper found that over 50 percent of the samples, pulled randomly from the non-concentration response selected batches, produced an effect on at least one hormone.

The ability of the prescreen to miss these potentially endocrine-active chemicals was considered unacceptable, even from a screening

perspective. The goal of a screen is to cast a wide net with an eye on setting priorities.

And I'm going to move on. While the Agency gave information about the pathway method, compared to the validation efforts of the low-throughput assay, it did not give information on how the new test performed compared to Tier 1 and List 1 tests. This was considered a limitation in the ability to assess performance.

An additional comment was made that current tests were considered limited as they do not adequately characterize activities of phthalates. These chemicals are known to interrupt the steroidogenesis pathway. And that is the extent of the comments that I have at this point.

DR. JAMES MCMANAMAN: Thank you. Dr. Clewell.

DR. REBECCA CLEWELL: I think I agree with many of those points. We have some difference of opinion in this group, but I think all of the opinions are valid and should be considered. The conclusion of whether or not it's replaceable, at this point, is not unanimous, I would say, probably. Whether the low throughput could be replaced with the high throughput,

whether that's feasible right now, I don't know that we're all on the same page with that one.

DR. KATIE PAUL FRIEDMAN: Okay. I thought you said something else.

DR. REBECCA CLEWELL: It's not unanimous is what I was saying. And the reason I'm saying that is because I think that it would be reasonable to go forward with replacement of the low throughput with the high throughput. I do think that I would put some caveats with that. I would like to see more robust evaluation of the cytotoxicity measure.

And honestly, it would be possible to do a direct evaluation of mitochondrial function in the same plates as the steroidogenesis. And I would say that would be something that would be a very high priority, given all of the scientific expertise you've heard over the last two days with the knowledge that it's tremendously important to have significant mitochondrial function to get steroidogenesis.

That and the prescreening method -- and we went around about that yesterday, so I won't belabor it. It's just making it clear that what was done for one particular purpose, in a research and

development situation, is not necessarily the same implementation that would be taken with the EDSP.

I think there's a lot of people with strong misgivings about doing a high-dose prescreen, and eliminating chemicals from that in a EDSP evaluation-type effort. And so, we would certainly want a multipoint-dose response for this type of a situation. I guess that's it.

I just wanted to put it on the record that I don't think that the assay is not scientifically sound alternative for the low throughput. I do think that it's a sound alternative, but I do think that there are significant analyses, and maybe a few more validation experiments, that should be run if it's going to be used in a screening type effort.

It wouldn't really be that big of a work. I think the goal of NCCT is a good one and is to move quickly through a lot of chemicals. And that's partly the goal of EDSP. I get that too. But sometimes there is real value in taking a minute -- in taking a little bit of time to really validate your assay before you put it through thousands of chemicals.

A few more validation studies to look at multiple replicates, multiple batches, different types of statistical analyses, it would be valuable. It would tremendously shore up the case for the high-throughput version of the assay.

DR. JAMES MCMANAMAN: Thank you. Next up is Dr. Jett. David, are you on line?

DR. DAVID JETT: Yeah. Can you hear me?

DR. JAMES MCMANAMAN: Yup.

DR. DAVID JETT: Hello?

DR. JAMES MCMANAMAN: Yeah. We can hear you David. Can you hear us?

DR. DAVID JETT: Yes. I can, sorry about that. I'm sorry I couldn't be there today. I was there a little bit yesterday, but I'm home with a pretty bad head cold and that's why I'm not there. I had this bad dream of the CDC identifying me as a ground zero for this weird outbreak of illness among our nation's top endocrine disruptor scientists, so I decided to stay home.

I guess I just have a few comments. First, we have a highly-qualified group of ad hoc members with far more expertise that I do in this

area, so I defer to their comments. Especially about the more detailed aspects of the assay.

My first comment would be this issue of fit for purpose, I'm very supportive of that. That is, really high-quality, rigorous and reproducible methodology, but a good match with available resources. I was a little concern about -- there was a comment in the whitepaper about, I think, including (inaudible) studies and aromatase assays and other things that were sort of a barrier within the prioritization process.

And I was just wondering whether these would be more appropriate for follow-along studies after initial screening. That was one area; and it may just be an misunderstanding of the process. That was the first one.

The second one out of the three was the pre-stimulation with forskolin. Admittedly, again without a full understanding of the steroidogenesis assay, the question that I had was whether this pre-stimulation affects sort of the dynamic range of the assay and its ability to detect chemical that stimulate rather than inhibit.

Again, this may have a simple answer from those that are more knowledgeable. For example, the dynamic range for up or down may be retained with this particular level of stimulation. Or perhaps it could be tweaked in that regard. The actual level of stimulation, if that's possible.

I do think EPA is aware of this. I saw a statement, I think on page 104 is what I have here, where they talked about this issue of not being able to detect. I think it was with a couple of chemicals that may have been less sensitive to E2 increases due to pre-stimulation with forskolin. I take that as they're aware of that.

And then the final one was the part about where significant affects were observed for a given hormone when two consecutive concentrations demonstrate a significant affect. And I was just wondering if there's been any thought as to whether the concentrations are far enough apart for this to be meaningful.

For example, you could have two only very high concentrations showing activity. I just thought maybe the Agency should be confident that this approach does not sort of undermine the whole purpose

of the multi-concentration approach. That's all I have. I'll send my comments through the email.

DR. JAMES MCMANAMAN: Thank you, David. Dr. Nagel.

DR. SUSAN NAGEL: I apologize, I think I thought I understood the question. And now it keeps going around and around a little bit. I just wanted to follow up on those two comments. Specifically, comment on the suitability of the assay as an alternative.

As is today is a different answer as the suitability assay in general. What I hear from EPA is that this is an ongoing process to optimize the assay. And we're giving tons of feedback today about how to do that. Yes, I think it is absolutely a good alternative, the high-throughput assay is.

Today am I convinced where it's primetime? No. I don't think it's primetime. Once again, I apologize if I am not still quite crystal on the question.

DR. JAMES MCMANAMAN: Thank you. Dr. Pullen Fedinick.

DR. KRISTI PULLEN FEDINICK: I was looking through my notes, I think that everything may

have been covered that I wanted to discuss. That's great. I think just making sure that the test, again, should be more sensitive than specific. And so, I think for this particular comparison it's really important.

Especially for the T -- I think it was a T down with 55 percent or 67 percent depending on whether or not some of the chemicals were removed. I think it's really important from a public health perspective to ensure that the tests are sensitive. I think that everything else was covered. Thank you.

DR. JAMES MCMANAMAN: Thank you. Okay, I'll open it up to other panel members, this charge question. Any comments? Dr. Pullen Fedinick.

DR. KRISTI PULLEN FEDINICK: Sorry I forgot. There is one that I think wasn't mentioned. Did we say the Tier 1, List 1 chemicals? I don't remember if we said that or not. I'll just repeat it just for the sake of repeating it.

Being able to have the ability to look across those different tests to compare the high-throughput test to the List 1, Tier 1 test results, would be really important in terms of being able to evaluate these tests in their real-world applications.

DR. JAMES MCMANAMAN: Okay, thank you. Dr. Clewell.

DR. REBECCA CLEWELL: I actually want to respond to one of the comments because I was actually watching Dr. Paul when the question was made. Because I thought maybe there was some misunderstanding. And maybe if we bring her up she could respond after, like we did before.

The stimulation with the forskolin used as a positive control. The question yesterday was that compared to DMSO alone or to forskolin pretreated control. It was to a forskolin pretreated control. Okay, so she's nodding her head. I wanted to make that clear because I think that question happened yesterday, and I sort of watched the miscommunication happen. But I wasn't sure if I was right. But now I think I am.

I think that that question of whether the forskolin, as a positive control, is allowable, I would say that it probably is and that was a minor miscommunication that happened yesterday, I think. Because it's on the record now that it was done one way, and if it was done the other way, I feel like it's important to put that on the record.

DR. JAMES MCMANAMAN: Okay, thank you. Other comments. Okay, seeing none. Back to the Agency.

DR. SEEMA SCHAPPELLE: I don't have anything to add at this time. But I will offer to Dr. Bever or Dr. Paul-Friedman if they'd like to add anything.

DR. JAMES MCMANAMAN: My question is whether the recommendations and the comments were clear?

DR. RONNIE JOE BEVER: Yes. They were clear to me. I thank you for your input. I will pass it to Dr. Paul-Friedman.

DR. KATIE PAUL FRIEDMAN: Just to respond to Dr. Clewell and the comment made about forskolin pretreatment and then forskolin used as a positive control. Dr. Clewell is correct. We did forskolin pre-stimulate washout, and then forskolin was treated just like a test chemical. And so, the comparison is back to the forskolin pretreatment. That's how that work was done. It was treated essentially like an experimental test chemical, like any other chemical in the set.

I wanted to provide that clarification on the record. I think everything else was really clear, and I appreciate the comments that were made.

DR. JAMES MCMANAMAN: Okay, thank you. Well, that ends this charge question. We'll move on to the third.

DR. RONNIE JOE BEVER: Question 3; please comment on the strengths and limitations of integrating multiple hormone responses beyond testosterone and estradiol -- i.e., using 11 hormones versus 2 hormones -- in a pathway-based analysis of the high-throughput H295R assay. Please comment on the suitability of this high-throughput H295R pathway model, using 11 hormones, to serve as an alternative to the low-throughput H295R assay. See Section 3.7.2.

DR. JAMES MCMANAMAN: That's not the charge question that I have listed in front of me. I see it's on the board up there. Does the panel have -- it's the same one? Oh, in the handout? Okay. Good. That's important. I'm hoping that the member assignment is correct. Dr. Androulakis, Dr. Clewell, Dr. Ehrich and Dr. Nagel are on this charge question. And Dr. Androulakis is the lead.

DR. IOANNIS ANDROULAKIS: Thank you. Once again, what I'll do is I'll summarize our thoughts and then my colleagues will correct anything that I said that is wrong. One comment that I would like to make, if I may, this was kind of an interesting question, because literally, 3 sits between 2 and 4. Because for us to answer 3, first of all we need to have a good assay, which is really what 2 was. And then for that to make sense, you need to have a good statistical method, which is what 4 is.

The reason why I'm saying this is because you will hear certain things that you've already heard in the discussion of Charge Question 2. And I'm going to guess that we may mention certain things that you'll hear again when we discuss Charge Question 4.

In terms of the strengths, we feel that the high-throughput assay does monitor obviously several hormones encompassing a simplified network of cross-regulatory element along the steroidogenesis pathway. As such, the advantage that it offers is that it enables more of an integrated response as opposed to the isolated elements, E2 and T.

The analytical system has the potential of offering higher sensitivity, since not one has the options or the opportunity. Again, assuming that all the issues with the assay have been resolved, so now you can actually measure coordinated responses as opposed to more isolated elements along the pathway.

The 11 hormones presented for distinct classes; so that was felt that it does add significant diversity to the measurement. It's felt that the ability to measure these multiple elements has the potential of not only proving the accuracy of predictions, but a may be -- and this is something that was also mentioned yesterday -- to provide some additional mechanistic insight and information, and maybe mode of action or things like that.

The two assays use the same cell line. So, the assumption is that, okay, whatever we learn from one assay can be transferred to the next. The network of the 11 hormones basically builds on the already existing two one, so it's an augmentation of the future vector.

The ability to measure these multiple hormones, and the complex part that emerge -- and I'll come back to that in a minute -- they really

demonstrated that exposure to chemicals is far more complex. And that most likely provides further evidence as to why one should be moving towards these more high-throughput and integrated pathway-based approaches.

It's felt that pretty much the high-throughput assay performs comparably to the low-throughput assay. And I'm primarily referring to the confusion matrices when E2 and T were measured, especially after the revised metrics and so on.

Again, even though this will probably be discussed in a minute in Question 4, but we also feel that the development of the distance metric was important. Because it's not enough to generate 11 dimensional vectors, you have to be able to do something with them. You also need to have the ability to analyze and really process your data.

We also believe -- and again, that's something that is also mentioned in the whitepaper -- that in the long run, and again along the idea of moving more towards hopefully this dynamic and kinetic model, the ability to measure things at a pathway level, even if the pathway is sort of somewhat arbitrarily constructed, will definitely provide

critical information for moving towards more qualitative systems toxicology models.

However, the group felt that there are a few limitations that sort of need to be addressed. And as you will see, most of them really relate to the assay. It's felt that the way it is implemented -- the high-throughput assay -- and I believe that was also mentioned by EPA yesterday -- has lost some of the advantages, such as the low-throughput assay has been validated across multiple laboratories, there are fewer technical and biological replicates.

Despite the fact that the confusion matrices indicated some strong correlation between low and high-throughput assessment, it was not exactly clear -- at least based on the results that were provided -- whether these trends would remain valid as one starts moving towards the integrated metric as expressed through the Mahalanobis distance.

And especially when you move to chemicals that actually seem to be activating surer hormones. Now all of a sudden you have a future vector that has a few dominating components, but then maybe an equal amount of hormones that basically exhibit lack of response.

One of the points that was also mentioned in the public comments, is that on one hand the Venn diagrams were very informative; and informative in the sense that they demonstrate that what we observed is really a complex response that needs to be thoroughly represented. But at the same time, interpreting this result was not clear.

Just a few numbers if we look at the data, there is a least 400 chemicals that impacted only 1 or 2 hormones. There's 300 chemicals hitting between three and five. Then there is about 307 chemicals that hit all 4 pathways. It seems as if a lot of things are becoming very promiscuous. The question is, it's good to have this information, but then it's not exactly clear, or it's not discussed very thoroughly how one would expect to sort of manage that kind of information.

Another point that was brought up is that yes, measuring these 11 hormones, especially in the context of a pathway, is extremely valuable. But however, to provide support about the clear need for the added hormone measurements, it would have been nice if there is some kind of a comparison of how many chemicals would be called a hit.

If one looks at, let's say, E2 and T, versus if one were to look at a combined score. I guess, in other words a way to rephrase this comment is, are significantly more chemicals identified when the additional hormones are measured? That's, I guess, is the simple way to phrase that point.

The whole discussion we also had this morning, the application of the cutoff of the at least three hormones being changed. It's something that probably has to be revisited. Again, it was emphasized -- also in the morning, but also yesterday -- that a lot of that was kind of a decision that was based on adding the resources in developing this high-throughput assay.

However, an interesting question would be, what are the results that basically indicate that hormones should become activated? What does that really mean? Is there anything there that maybe we have missed?

The Mahalanobis distance, again, I realize that it's the charge of question 4, but in this context, we believe, it does become important. There was a lot of discussion about what happens when

you have a lot of hormones that are active, and then you basically have a more dense vector.

But still, the question is what happens when you have like weakly-active chemicals. What happens if you hit one or two or even three of these hormones? I guess the point here is a little bit more complicated because one would have to probably look at this data, because otherwise the question is, is this analysis, the way it's been presented so far, sort of bias more towards chemicals that appear to have a more across the pathway affect, as opposed to chemicals that maybe they hit some targets upstream, but then they don't propagate.

The question is, let's make sure that the analysis based on the combined score is not sort of bias towards things that appear to have a more profound and broad affect. I guess the committee's question was whether this has not been done, whether the suggestion is it should be done; and if it has been done, we think that this information is worth sharing.

One of the things that was sort of also brought up is that, yes, a lot of hormones are measured. However, it was felt that there is a lack

of reference in terms of the additional components that are being added.

The way we understand it, is there is some information about at least the two hormones of the low-throughput assay, but there is no positives or negatives for the remaining hormones. Having this would actually be important.

There was some mention, and we talked a lot earlier also today about this whole issue with the DMSO, water, whatever. But as a group we kind of felt that this is not something that is special to the high throughput. It's probably something that could be discussed, but not in this context.

As was discussed earlier, the whole issue of cell viability -- and again, also earlier there were discussions about if they are maybe analyzing the results and introducing either the Z-score, maybe finding a different viability set test and so on. But we feel like at the end of the day the question is whether the, sort of, limiting of the 70 percent viability, does that skew the results? Do you feel there is evidence that it does?

We felt it would also be interesting to see what would happen if the viability cutoff were to

increase. In other words, the question would be how sensitive is the method to the selection of some of the parameters that are used for determining the kinds or types of data that's revived?

Increasing the dimensionality of the feature vector, going from 2 to 11, definitely moves in the right direction in the sense that it does increase the information content. But it does make the interpretation a lot more difficult. That's where in charge question 4 comes into the picture. And basically, the point being that it's nice to be moving in the higher dimension, but then at the same time let's make sure that everything else sort of keeps up.

To summarize, overall, we felt that the high-throughput assay offers significant advantages over the low-throughput assay. Measuring the multiple hormones in conjunction with the development of this integrative statistical method is important. It's definitely moving in the right direction.

We feel that the high-throughput test can and should serve as an alternative once everything, of course, has been realized. However, again, much like the group before, although we feel that this is a scientifically sound alternative -- the

high-throughput assay to the low-throughput assay -- there are specific assay condition questions. Viability for instance was one of them, as well as analysis methods, multiple hormone effect, cut off, the scores, what happens when you have low-hit count and so on, that have to be addressed before an implementation takes place.

Also, the issue of generating data when positive or negative controls do not exist. How do we assess and evaluate that? We feel that this could be very important for prioritization or hazard ID. That's pretty much our summary. Thank you.

DR. JAMES MCMANAMAN: Thank you. Dr. Clewell.

DR. REBECCA CLEWELL: I think Ioannis did a good job of summarizing because there were a lot of comments he had to summarize. I just sort of pulled numbers from a plot. They're not going to be totally perfect numbers.

When I looked at that plot that we went over several times yesterday, at least 400 chemicals only changed 1 or 2 hormones. Of approximately 1,600 to 2,000 chemicals, about 400 of them changed 1 to 2

hormones. About 300 hit 3 to 5 hormones. And about 300 chemicals hit 5 or more hormones.

And then when we look at the Venn diagram, we see that 307 chemicals hit all four pathways. The conclusion from that was that most chemicals hit most pathway. But that's actually not the case. That's less than a third of the chemicals that hit most of the pathways.

That's not most chemicals, that's one-third of the active hits. And even less of a percentage of the overall chemicals that were tested. I think that that's being overstated. I think it's worthwhile to look at, but I think it needs to be looked at in a couple different ways; to say whether -- if we want to make the claim that a rollup of the pathway is a more sensitive way to look at a steroidogenic effect, than the analysis needs to be done in some additional ways.

And what I would like to see is something where we actually look at what if we used one or more chemicals. What if we used two or more chemicals? What if we used three or more. I mean, just a systematic evaluation of what happens to the numbers when we do that.

Ioannis mentioned this a little bit, but I wanted to highlight it. I worry that what happens if we only look at the chemicals that affect a lot of -- and I could see why you'd say oh it's more important if they affect a lot of the hormones versus one hormone. But not necessarily. And we bias towards the upstream of the pathway.

Everything is dependent upon pregnenolone and progesterone. If we hit one of those guys, everything else is going to fall out. But that doesn't mean that something that specifically hits cortisol or specifically hits estrogen, is not a concern. What I'd like to see is a bit more analysis around that. And a bit more analysis around the Mahalanobis score -- which I might be saying wrong -- just to demonstrate that it would work as well for weakly-active chemicals that only hit one or two hormones, as it does for weakly-active chemicals that hit many hormones.

You may have a gut feeling that it does that. And you may have done those analyses. And if that's the case, wonderful; and then I just think it maybe needs to be added to the record. But I wasn't able to find that information or I missed it. And I

think that's important. Because honestly, from a public-health standpoint, if it hits progesterone or if it hits estrogen, either way it's a concern.

And that's all just on the analysis side of things that I think would really shore up the use of this. Because really intuitively it makes sense to try to use a more comprehensive steroid-pathway approach. I like the approach and I think some simple analyses would make it a much stronger argument.

One of the comments that was said this morning -- I don't remember who -- said something about a 30-percent hit rate. And that does seem to be about right. And if that's the case, that's pretty high.

We have done, in my lab, an evaluation of the hit rate for ToxCast assays overall. The average hit rate -- at least from the 2015 release of data -- is 14.7 percent. The average assay hits for about 14.7 percent of the chemicals. And only about 16 percent of all the assays in ToxCast have more than a 30-percent hit rate.

It has to be a pretty permissive assay to get a 30-percent hit rate, is what it appears from

our analysis. If you're getting a 30-percent hit rate for steroidogenesis, the conclusion is that steroidogenesis is one of the more common -- the logical kind of conclusion would be that steroidogenesis is one of the more common chemical pathways. And I don't think the in vivo data bears that out.

All of this is sort of an intuitive -- I haven't run all of the numbers, but some of the numbers. And I think it's worth going back and looking at that and saying so why are we getting a 30-percent hit rate. Are we getting it because -- if that 30 percent is correct. Because I have to mention the cytotoxicity thing.

I wanted to specifically say, from my own experience, I have run a lot of steroidogenesis assays in Leydig cells, rat Leydig cells. Never the H295R. And in that case, as little as 10 percent reduction in ATP consistently correlated with a drop in steroidogenesis whether I was using a positive or a negative control.

Even in negative controls, in known negative controls, a 10 percent ATP drop was a cutoff for me. Because even in negative controls, I would

see a reduction in testosterone. And I wasn't doing an 11-hormone panel, but I think that's important. ATP measures mitochondrial health; and those two are just so intricately related that, once again, we come back to we really need a measure of mitochondrial health. And we need to do a more careful evaluation of a cutoff there.

And then finally, the recommendation -- we don't have positive controls for the glucocorticoid pathway or for progestins. And I think it's valuable to have these as a prioritization tool where I think there is some caution to be had. And what I would like to see is a recommendation for how do we use something like that as a prioritization tool. And not accidentally trip into a hazard ID.

If we see something like this where we don't have positive controls, but the overall rollup of the pathway is so consistent with previous steroidogenesis assays, that we say it's good enough. We have the numbers, we're going to go forward with our prioritization. But we also have, on the record, inhibition of cortisol, or increase in cortisol. And how do we refrain, or prohibit, a sort of preemptive or early -- premature is the word I'm looking for --

hazard identification in that way. I end with that. I will turn off my microphone.

DR. JAMES MCMANAMAN: Thank you. Dr. Ehrich.

DR. MARION EHRICH: I think I have to compliment the EPA. This is pretty complicated. Reading this Karmaus paper, it sounds like you could do LC-MS/MS with the same well so you could get all of those from one sample, which is pretty impressive. And that is a very sensitive method.

One always has to be aware that when you're measuring levels, it doesn't necessarily mean functional change, but it's the best you have. I think it's good that you're doing that. And you've already recognized that you have only limited capability for handling metabolites and so forth. But the system has a lot of cytochrome P450s in it, so it's better than many. That was with the low-throughput assay, but I think you've chosen well to use the same cell system that was being used there.

Basically, that's pretty much what I have to say. But I think this is a step forward for sure.

DR. JAMES MCMANAMAN: Dr. Nagel.

DR. SUSAN NAGEL: I think the only thing I have to add is just -- and you probably have this. But I didn't think to ask for it earlier, of just a vary simplistic, direct comparison of relative sensitivity of the assay. So, for IC50s and AC50s for the chemicals between the two assays. And you probably have that, but I would love to see it.

And most definitely making that really transparent. Because when you jump to the extremely complex integration of the data, with the model, and in fact, just a simplistic view of it would be, I think, extremely important for sensitivity.

DR. JAMES MCMANAMAN: Okay. This charge question is open to the rest of the panel. Anyone has comments? Yes, Dr. Pullen Fedinick.

DR. KRISTI PULLEN FEDINICK: I had a question, actually, for the folks who were on this. Were you saying that the assay has potential to be used as an alternative for the low-throughput assay, but it's not there yet? If there was a single line to answer this question.

DR. SUSAN NAGEL: That is correct. That is what I'm saying.

DR. JAMES MCMANAMAN: Okay. That was Dr. Nagel.

DR. REBECCA CLEWELL: I would say that yes, it has great potential to replace and even be better than the low throughput. Particularly, if you can involve the multiple hormones. I think there's real advantage there. And I do want to highlight that, because I just said some fairly tough comments, I guess.

I would say there's real value and I don't think it's far away from being replaceable. And it's mostly all on the analysis end to shore up confidence.

DR. JAMES MCMANAMAN: All right, thank you. Other comments? All right, then I think we'll move on to the next charge -- oh, I'm sorry. Go back to the Agency.

DR. SEEMA SCHAPPELLE: Nothing from me.

DR. KATIE PAUL FRIEDMAN: I just wanted to respond to Dr. Nagel's comment about potency comparison. There is a really simplistic table that does that. And admittedly, there are quite a few supplemental files to this chapter. I believe it's supplemental file 10, is a table where it actually

looks at the hit call in low throughput and high throughput. And then the LOEC, as reported by the OECD intra-laboratory validation versus the MTC concentration from high throughput.

And then there's some notes. There's a table that makes that comparison. Not very sophisticated, but very simple laying out chemical by chemical.

DR. JAMES MCMANAMAN: Okay, thank you. If everything's clear, we'll move on to the next charge question. Charge Question 4.

DR. RONNIE JOE BEVER: Question 4; the work herein presents a novel statistical integration of multiple hormone responses indicative of steroid biosynthesis in the high-throughput H295R assay. A summary statistical metric, the maximum mean Mahalanobis distance, has been suggested as a tool for use in prioritization of chemicals.

In addition to the use of the maximum mean Mahalanobis distance to indicate the magnitude of potential effects on the steroid biosynthesis pathway expressed in H295R cells, an examination of the hormone responses that contribute, to the maximum mean Mahalanobis distance, may provide valuable biological

information to inform the weight of evidence evaluations performed for chemicals subjected to the EDSP Tier 1 evaluation.

Please comment on the strengths and limitations of using the maximum mean Mahalanobis distance. And the pattern of steroid hormone responses in the high-throughput H295R assay for chemical prioritization and weight of evidence applications. See section 3.2.4, 3.3.2 and 3.7.2.

DR. JAMES MCMANAMAN: Thank you. The members of this charge question are Drs. Berrocal, Androulakis, Barr, Pennell and Weller. Dr. Berrocal is lead. You're up Veronica.

DR. VERONICA BERROCAL: Yes. I will read the summary of the comments I received from some of the panel members that were assigned to this question. I would say oftentimes members of the panel, but I'll just refer to the panel that responded to this question and sent me comments.

Members of the panel believe that the efforts of the Agency, to consider multiple hormone responses simultaneously, is a great attempt at obtaining an integrated and comprehensive indication

of the magnitude of the potential effect of a chemical on the steroidogenesis pathway.

In reviewing the proposed maximum mean Mahalanobis distance approach, as a tool for prioritization of chemicals, we have identified the following strengths and limitations. I will start with the strengths.

The proposed approach for assessing steroid biosynthesis generates multidimensional data, 11 hormone responses for each chemical of various concentration. The maximum mean Mahalanobis distance is a way to summarize this multidimensional data into single scale or quantity.

For as exotic as it might seem to non-statistically trained individuals, the metrics proposed by EPA has close ties to quantities used in statistics, such as Hotelling's T squared test statistics. To test whether there are significant differences between two groups, one looking at multidimensional data. And this is also mentioned in the whitepaper.

More loosely, the mean Mahalanobis distance can be thought as the multidimensional equivalent of the disease score derivation that is

done for a univariate normally-distributed observation. And so, it can be used to flag outliers.

An advantage of using these metrics is that it allows to combine measurements of multiple hormone responses into a single summary measure, accounting for the second moment of the sampling distribution; which means accounting for the viability of each individual hormone response measurement, as well as the correlation among the various hormone responses measurement.

Working with such a metric allows to control for high-variable hormone responses. And also allows a way to deal with the multiple testing adjustment that one would have to use if tests for each hormone were conducted separately.

It has been brought up that the Mahalanobis distance is an appropriate outlier detection only for multivariate normal data. And concerns are being raised about whether the data considered here is normally distributed. While some panel members believe that this might be a concern potentially of marginal importance, I personally believe that the type of data considered here has been already assessed to be normal as it's mentioned in the

paper by Zhang, Chung and Oldenburg, which is referenced in the whitepaper.

Additionally, it is important to notice that the mean Mahalanobis distance, in the proposed framework, is used on the log hormones response and not on the raw data, which potentially would have a skewed distribution if it's told in the univariate sense.

It is also important to notice that the whitepaper indicates prioritization of chemical based on the maximum mean Mahalanobis distance, over concentration, which would yield a conservative approach for flagging a chemical as an outlier with respect to the control. Those are the strengths.

In terms of limitations, it is unclear what type of effect of a chemical on a steroidogenesis pathway would the proposed maximum mean Mahalanobis distance metric approach tend to flag. It is the intuition of the panel members that this approach would tend to flag mostly chemicals that deviate from the expected relationship between hormone responses. And it would not allow to prioritize chemicals that displays absolute different from control regardless of the sampling distribution of the residual. It would

be great to investigate if such statements are correct.

More specifically, panel members believe that the maximum mean Mahalanobis distance metric might result in prioritizing a chemical that has relatively small absolute difference from control in any single hormone, but unusual combination of hormone responses with respect to the sampling distribution of the residuals, above another chemical which would have very large deviation from control, but which fall closer when adjusting for typical correlation structure.

Another example would be the following; two hormones could be highly correlating in a positive direction. If one of the hormone levels is above control level at certain dose than the other, it would be expected to be also above control level at that dose for the other hormone. If instead, the other hormone is below the control at that dose, the Mahalanobis distance metric could be large, even if the individual levels of the two hormones aren't very different from control.

Another issue is the identification of critical values in Type 1 error rate. The approach

uses critical values from a method developed by Nakamura Imada in 2005, which requires equal sample sizes across comparison and non-covariance matrix. Neither of which is actually the case in the analysis presented in the whitepaper.

As mentioned in the whitepaper, nominal Type 1 error rate will not be achieved. The whitepaper implies that a Type 1 error rate is approximated under this approach. However, without any numerical result to support this assertion, you have to know how close this approximation is.

Some panel members have the following suggestions. Perform extensive stimulation studies evaluating the Type 1 error rate of the proposed method, using the data in the report as motivation for the simulation setting. If simulation studies have already been performed, then they should be cited. Such studies are vital if this methodology is going to be a standard methodology going forward.

It's not clear also why 1 percent Type 1 error rate was used instead of the more conventional 5 percent Type 1 error rate. Was this used because of concern of an inflation of the Type 1 error rate? Simulation studies can also help determine whether

this is an adequate correction, or if it is too conservative.

Another issue is the estimation of the covariance matrix. The mean Mahalanobis distance is dependent on the knowledge of the covariance matrix. From the description in the whitepaper, it appears that all data that was now removed, for several reasons, was used for estimation of the sample covariance matrix. Regardless of whether the data was relative to a control chemical or not, regardless of the mode of action of a chemical, and regardless of the concentration value.

This is a statement that I'm making as a statistician who doesn't understand, very well, the biology. It might be plausible from a biological point of view that the correlation between hormone responses would differ depending on the type of chemical, control versus chemical, and the concentration level.

The whitepaper, on page 78, suggests that all the available data was used, and that this will ensure a large sample size and that the estimation of the covariance matrix is precise. I personally am not sure about whether this statement is

valid, given what I just stated in terms of the concentration and the type of mode of action of a chemical having an impact on the correlation between hormone response.

My fear is that the estimates of the viability in the hormone response might be overly inflated, and that the estimates of the correlation might be attenuated. And that could have consequence in terms of an inflation of the Type 2 error rate. I think it's important that there are efforts to investigate how to more efficiently estimate this sample covariance matrix.

There are also other specific comments that have been raised in the whitepaper regarding specific details. For example, how was the values below a limit of detection handled. Or the reason that two hormones were excluded. Another issue was the critical value and the critical limit that were used interchangeably in the paper. And some other issue regarding the calculation of the confidence interval.

In summary, and this is, again, a statement that I think summarizes the view of the people that sent me comments. Some panel members

explicatively stated that while this might not be an optimal statistical approach to integrate multiple hormone responses, due to some of the limitations that I just mentioned, this might not be the only approach as there are other methods to compute distances. For example, the Tukey distance, in which Dr. Weller can provide more information.

It should be recognized that this is a step in the right direction in the effort of developing a framework to assess chemical's potential for effect on steroidogenesis. And it is fit for purpose, which is what the EPA charge question is asking to comment on.

DR. JAMES MCMANAMAN: Thank you, Dr. Berrocal. Dr. Androulakis.

DR. IOANNIS ANDROULAKIS: I don't think I have anything to add.

DR. JAMES MCMANAMAN: Dr. Barr.

DR. DANA BARR: I have nothing to add either.

DR. JAMES MCMANAMAN: Dr. Pennell.

DR. MICHAEL PENNELL: Most of my comments were captured there in that summary. And overall, I agree that this is a step in the right

direction. But since you asked for limitation, there's one additional one I did have. Not saying I have a solution, but it is a limitation. The fact that this methodology is just a very general methodology.

You're using it as a method to flag pathway affects. Because you're giving it that type of name, it seems like it would be preferable -- it would be a methodology which actually incorporates knowledge about the pathway, rather than just using some generic method for multi-variate data. I think that's it.

DR. JAMES MCMANAMAN: Thank you. Dr. Weller.

DR. GRANT WELLER: All my comments have been captured primarily by Drs. Berrocal and Pennell. Just to follow up on a reference that Dr. Berrocal made. There are other alternatives for calculating sort of distance of outlying this in multi-variate data. The one that Dr. Berrocal referenced is the Tukey's half-space depth, which I can add a reference too in the written summary.

But for the purpose of sort of ordering, or determining which chemicals are sort of

most outlying or most extreme, it might be an alternative to at least benchmark or compare the results to the existing method, which is the Mahalanobis distance.

DR. JAMES MCMANAMAN: Thank you. I guess we can open it up to other members of the panel now.

DR. KRISTI PULLEN FEDINICK: Sorry for so many questions. I just had a clarifying question. When you say it's fit for purpose, what did you mean by that? I didn't quite understand that.

DR. VERONICA BERROCAL: This is actually a comment that I think resonated with what Ioannis said. Maybe he can address.

DR. IOANNIS ANDROULAKIS: When we discussed it, I guess, the point here is not assess or evaluate that if you can do it in different ways, try all of them and pick the best one. The way I sort of interpret that, is that there's a question. The question is you have this pathway representation; and what you really try to do, is you really try to rank things. And this is one way of ranking.

I think the assessment should be, you know, as Veronica sort of outlined, what are the

limitation in terms of providing a somewhat correct or represented ordering. And not so much as to whether you can use, you know, five other different metrics.

That's what I meant. It's fit for purpose is a specific question, at least that's how I understand it. This is one way of answering the question. Therefore, it's fit for that purpose.

DR. KRISTI PULLEN FEDINICK: If I can just clarify that too. You're saying essentially what was said for the last one, that there is a lot of potential in these methods, but there's still progress that needs to be made before they can be implemented for prioritization in weight of evidence applications? Just to -- is that -- Okay.

DR. JAMES MCMANAMAN: Thank you. I have a question for Dr. Pennell, or for the panel for that matter. You mentioned that it would be better if you could adapt a method that included information about the pathways. Does such a thing exist in any form?

DR. MICHAEL PENNELL: Well, I mean, I'm not saying I have the methodology available. But I'm just thinking with dose-response models, right, people

tend to prefer dose-response models that sort of reflect the mechanism of the toxin or the disease, right. And I feel like maybe it's the label of the type of effect which kind of -- calling it a pathway effect. I know you're trying to find an effect on the 11 steroids as a whole. Kind of infers to me that this is some sort of method that is incorporating knowledge about that pathway.

Again, I'm not saying that -- I don't know how easy would it be to do this, but that is a limitation. It's less ideal because this is not something that I feel is particularly tailored to this problem.

DR. JAMES MCMANAMAN: Is there a modeling approach that could be used? It looks like Dr. Berrocal has a --

DR. VERONICA BERROCAL: No. I was wondering if actually when you made that statement, you were mentioning more about having a covariance metrics that somewhat is reflective of the known or the expected relationship between these hormone responses, rather than having it estimated from data. I don't know if that's what you were trying to get at.

DR. MICHAEL PENNELL: That's not exactly what I was trying to get at. Again, I'm not -- I do have a Bachelors in Biology, but I'm not a biologist. I'm thinking there's a pathway, right. There's relationships between these different steroids, right. And that sort of confirms some sort of directional relationships, maybe, between these steroids, right. If you have an effect on one, that might infer an effect on another. Is it necessary to consider them sort of almost equally weighted in this multi-variant analysis? Or if you hit one part of the pathway, is that good enough to find an effect because you naturally expect, sort of, downstream components to also be affected. That's kind of what I had in mind.

DR. JAMES MCMANAMAN: Dr. Androulakis.

DR. IOANNIS ANDROULAKIS: I'll attempt to give an answer. I'd like to emphasize what I said before. I don't know if this is the right forum to have this discussion, because the question here is whether this method can help analyze the data.

As a system biologist, what I would say is that maybe what one can look at is not -- as was mentioned before by Dr. Pennell. In fact, what we

have here, we have a graph, you have a network. And the question you're asking is how different are two networks? The networks have a structure which is the fact that two nodes, two hormones are connected. And each of the nodes have a value which is the level of the hormone.

I'm assuming maybe one of the directions would be that now I'm not just comparing the fact that they have 11 different values, but they have 11 different values which are placed on a network, on a graph. And then the question that I'm asking is, how different are these graphs. And that's where functional relationships, for example, between different hormones can come into the picture.

Now you don't look at two different numbers, because one can ask an interesting question, you know, if the beginning and the end of my pathway appear to show activity, but then nothing in between. And if there is a signal that basically supposed to connect the beginning and the end, if everything in between is dead then the two ends light up, I mean, what does that really mean. But again, that's a different question.

DR. JAMES MCMANAMAN: Going back to the non-biologist mathematician. Anybody who fits that description, is the math available? The modeling approach is available to handle this systems biology question, which it really is a systems biology question. If you understood that, could you adapt the Mahalanobis approach -- whatever it is? The question is, if you had that information, could it be adapted into a model?

DR. IOANNIS ANDROULAKIS: Most likely not that particular metric. But there are ways, completely different thing that one could do to basically address that. But again, I'll emphasize I'm not sure the overall discussion we had as far as the charge question is concerned. But the short answer to your question, yes.

DR. JAMES MCMANAMAN: So, it might be useful to include those references in a writeup about that as a possibility as an alternative. Yes, go ahead.

DR. MICHAEL PENNELL: First of all I'm not a mathematician. And statisticians would say I'm not even a statistician, because I'm a biostatistician.

DR. JAMES MCMANAMAN: And that, you're sure about.

DR. MICHAEL PENNELL: Yes. I'm sure about that. I do want to reemphasize that I do think this is a step in the right direction. And I do think this is a good analysis method. I was one of the members that raised the point that this metric can be influenced by, sort of, outliers that are not really caused by extreme values of the individual hormones.

More, it's sort of the relationship between the two hormones is not going the direction you would expect, which is not ideal. I don't know how often that would cause problems, who knows, but it's possible I guess. I do like the idea of summarizing sort of the effects on these 11 different hormones in one measure.

Because again, if you're just looking at individual hormones on their own, it's going to be hard to really determine whether there's -- in fact, just looking at the separate Dunnett's test, right, does affecting two mean there's an effect on this entire pathway? How about three? Who knows, right. I do like this approach.

DR. JAMES MCMANAMAN: Thank you. Other comments or questions? Marion.

DR. MARION EHRICH: We're looking at this out of the whitepaper, Figure 3.1. You're talking about directions that thing go. Can that be helpful for some of the statistical analysis?

DR. IOANNIS ANDROULAKIS: Yeah. That's basically what I mentioned before. That one way of looking at this is -- the way we look at it right now, it's just 11 numbers. Whereas the point is that these are 11 numbers that are placed on the network and each number belongs to each one of these boxes. And there are boxes that are directly linked, and there are boxes that are indirectly linked.

When you run an experiment and you compare two chemicals, you don't only have these 11 numbers, but you also have where these 11 numbers are relative to the network they belong to. Then the question becomes I don't just look at how these 11 numbers change, but I really look at how the whole thing, the structure, plus the pathway, plus -- or the network rather, plus the numbers have changed.

DR. JAMES MCMANAMAN: Okay, thank you. Other comments? Questions? Okay, hearing none, I'll go back to the Agency.

DR. SEEMA SCHAPPELLE: No need for clarification from my perspective. I will just thank the panel for the comments that we've received. I think this really provides us what we need for optimization of these efforts as we move forward in development of our orthogonal assays. And it really does speak to the nature of the program in terms of our kind of learning by doing design. And so, this does provide the feedback that we need to continue furthering our improvements as we go.

DR. KATIE PAUL FRIEDMAN: I thought all the comments were really clear. I wanted to add maybe just a clarifying point; another option for modeling is kinetic-based modeling. And in the whitepaper we do reference that there are existing kinetic models that would take pathway information into account.

The problem in implementing those, in this scenario, is that those kinetic models were not optimized for a high-throughput screening assay like the one that we used. And so, we would need to collect more data, time-course data, spend a lot more

time and try to see if we could be successful in developing kinetic models.

The other obstacles that I think is also mentioned in the whitepaper, is that typically those kinetic models in H295R -- there's a few papers -- they tend to be optimized using one chemical or maybe a couple. And of course, we have a many-chemical problem. If we took that route, we might encounter some challenges. I wanted to clarify that's an option that was considered in the whitepaper. And we didn't have the appropriate data to pursue it.

DR. JAMES MCMANAMAN: Okay, thank you. At this point I think we can take a break. We have two more charge questions left. Are the panelists who are up -- you ready Tom? Okay, so we have the option that we can try to get through these today and end it a little early. We want to take a 15-minute break and be back here at a quarter to 3:00, and see what we can get done.

[BREAK]

DR. JAMES MCMANAMAN: I think we can begin the last portion now. I think we're prepared to move onto Charge Question 5.

DR. RONNIE JOE BEVER: Charge question five. Please refer to white paper Section 4.2. EPA has identified adverse outcome pathways for thyroid hormone disruption related to potential xenobiotic-induced alterations of thyroid homeostasis.

Please comment on the completeness of the molecular initiating events, Table 4.1, key events, and adverse outcomes within the thyroid adverse outcome pathway network, Figure 4.1

Also, please provide information on any missing pathways, adverse outcomes, or other adverse outcome pathway related information, e.g. molecular initiating events or key events, critical for capturing the complexity of systems biology controlled by thyroid hormones.

DR. JAMES MCMANAMAN: Thank you. The panel members on this charge question are Dr. Zoeller, Belcher, Furlow, and Shaw. Dr. Zoeller is lead.

DR. THOMAS ZOELLER: I wanted to begin first by kind of reading what the white paper says about framing the issue here, and that is that EPA has previously demonstrated that estrogenic activity from nuclear hormone receptors, ER, and resultant cellular signaling pathways correlates -- or correctly predicts

over 85 percent of chemicals known to produce positive findings in estrogen-related in vivo assays, i.e. the uterotrophic assay. But thyroid hormone receptor activity, fails to predict the vast majority of thyroid hormone-related findings in in vivo studies.

This appears to be due to the high ligand specificity of TH receptors and the ability of chemicals to interact or act with different sensitivities on the multiplicities on the non-TH receptor elements within the thyroid pathways. As a result, a comprehensive pathway-based approach that incorporates screening for potential interaction with multiple MIEs is needed to effectively screen for thyroid disrupting chemicals.

Now, there are two important issues here. The first is linking MIEs through key events to an adverse outcome. And the second is identifying adverse outcomes that are specific to that AOP pathway initiated by a particular MIE. Clarifying this issue is critical in the strategy the Agency employs to achieve the goal of developing high-throughput assays that could populate an AOP providing information for thyroid disruption.

An example that the Agency uses is that of thyroid hormone receptor. The observation is that TR activity fails to predict the vast majority of thyroid hormone-related findings in in vivo studies. And the interpretation is that the ligand binding domain is too restricted. But the in vivo findings are serum T4, serum TSH, thyroid weight, and histopathology.

Even though there is ample evidence to support the conclusion that thyroid hormone receptor ligand binding domain is more restricted than that of ER, it's also true that a chemical could activate or inhibit a TR -- especially TR-alpha 1 -- without affecting these guideline endpoints of T4 and TSH, et cetera.

The point that I'm trying to make here is that the AOP, being investigated by these in vitro assays, needs to reflect the pathway that affects a specific adverse effect. To elaborate on that a little bit, we know the TR-alpha 1 doesn't regulate thyroid hormone levels or TSH, in humans or in rats or mice. Including TR-alpha in this calculation of this, let's say, balanced accuracy, is going to dilute the

findings because it's not relevant to the endpoint under investigation. That's a key issue to me.

The second thing is that if you look at all the chemicals that affect thyroid hormone levels in vivo, in guideline studies, and then determine in vitro what could be explaining that, that is going to require multiple MIEs or multiple high-throughput assays. But it's still going to be difficult for reasons that I'll talk about here, largely in terms of Table 4.1.

So, 4.1 is largely complete, I think, in terms of MIEs that control thyroid hormone action. Now the question talks about systems biology controlled by thyroid hormones. I'm not exactly sure if those two things are the same, but for the sake of my argument I'll assume that it is. Thyroid hormone action in tissue is what you mean by controlled by thyroid systems biology -- controlled by thyroid hormone.

One recommendation that I have for Table 4 -- and it's not so much for the table, but in terms of Agency thinking and planning as they move forward on this -- is that you add an additional column that includes the adverse outcome that would be

predicted to result from the interference with a particular MIE. The first thing that's going to do is require you to separate TR-alpha from TR-beta. And TR-beta 1 from TR-beta 2. We have enough information, both in humans and in animals, to be able to link what kind of adverse outcomes are going to be related to those particular MIEs. The same is true for really all the rest of them.

Many of the issues -- you're probably not going to have any evidence for linking the MIE to a specific adverse outcome. But that's actually important, because it tells you what we don't know and it's for you to be able to navigate, I think, especially at the beginning of this project. To be able to navigate, you're going to have to know what we know and what we don't know, and probably focus initially on what we know.

While this AOP concept is covered, to some extent, in Figure 4.1, to highlight it here would be this opportunity to, first of all, reference the scientific evidence for a specific MIE pathway -- related pathway. It's a complex system and the Agency has made great strides in organizing their work effectively. But articulating what we know and what

we don't know in this AOP kind of approach, is going to be important.

Now, this is kind of a trivial point, but since the recognition of cellular transport by Gruters (phonetic) and others -- actually several years ago -- serum binding proteins have been called, officially, distributor proteins. If you call them transport proteins, it gets really confusing whether you're talking about serum transport or cellular transport. It's also now an antiquated kind of term. I think it would be better to call them distributor proteins. Serum binding proteins, I think, discriminates between transport proteins like MCT8. It would be useful to change that kind of language just to keep that clear.

For the hepatic nuclear receptors, two points here. The liver controls thyroid homeostasis to the same degree that thyroid synthesis and release does. Those are two ends of the regulation of circulating levels of thyroid hormone. If it's a normal physiological event controlling thyroid homeostasis, then it doesn't make any sense to consider it an indirect effect that's not relevant to the issues at hand.

Now, I think there's a lot of reasons to separate the nuclear hormones. In Table 4.1, it just has nuclear hormones. I think those should be separated also, because those are going to be separate MIEs that could be important to explain effects on thyroid homeostasis. The same is true for sulfation and glucuronidation.

The regulation of TRH synthesis, or neuronal activity, might also be important. TRH receptor assays are going to be important, but we know that there are mechanisms, both in humans and animals, that affect the TRH neuron that can be reflected in TRH gene expression. But it can also be reflected in other ways that could represent an MIE that's important.

As I said earlier, thyroid hormone receptors need to be separately identified. Also, the term TH transcription probably means TH regulated transcription. That could be updated. That's another kind of trivial point. This whole field of TH regulated transcription is incredibly large and complex. Biologically, but also in terms of system biology. This is going to be important to pay attention to.

The Agency makes the point that there are many MIEs that lead to a common downstream effect through key events, which are linked to a number of adverse outcomes that are species and life stage specific. This would be clearer if you did add another column that had those kinds of adverse outcomes also incorporated.

In the discussion of Tier 1 and Tier 2 of the EDSP, it might be useful to state that one is hazard identification and one is hazard characterization. Just to make it clear what the goal of those two things are as you think about how to use these high-throughput assays to replace something, even in Tier 1. Or even to prioritize what might go through Tier 1.

For Table 4.2, for Tier 1, as I said before, thyroid related endpoints that are captured in the pubertal assay are T4 and TSH, thyroid weight and thyroid histopathology. These are known to be separable in some cases. That is, some chemicals can cause a reduction in serum T4, both total and free, but they don't cause an increase in serum TSH. It's not clear at all how that happens, what the mechanism of that is, but it's not uncommon.

This means, first, that thyroid weight and histopathology are endpoints related to serum TSH, not T4 directly. It's important to keep that distinction clear, I think. Second, it means that in the absence of a clearer AOP that can discriminate between chemicals that have those two different effects, getting a balanced accuracy that is acceptable is going to be difficult, because those two things -- those two kinds of chemicals and their effects -- are going to dilute each other.

The Agency has identified that a 10 percent reduction in serum T4 isn't adverse outcome. This level of T4 reduction -- in fact, even an 80 percent reduction in serum T4 doesn't affect growth or body weight or brain weight. And therefore, it would be prudent if the Agency points out or stipulates that growth and body weight can be affected by low T4, but only under really severe circumstances. And that many adverse outcomes will occur that are thyroid hormone specific, while growth and body weight, as well as brain weight, are normal.

For Tier 2, thyroid endpoints are pretty much the same. It's a different exposure. Designed thyroid specific endpoints are the same, and

it's important, I think, to think about the neuro-histopathology, which could be predictive or reflective of thyroid specific disruption. Because not all neuro-histopathological endpoints are related to thyroid hormone.

Actually, it gets pretty complicated to think about thyroid hormone action in the developing brain because some things are not affected by thyroid hormones. So, it has to be very specific to thyroid disruption. In the investigation or in the identifying of those endpoints that are thyroid hormone specific, I think will give the Agency a lot of clarity on how to develop a model that's predictive of those specific kinds of endpoints.

Figure 4.1 complements Table 4.1 well, providing this kind of visual diagram of the various thyroid-related AOPs; but it's difficult to populate this figure with the resolution that I think the Agency really needs to employ this concept -- the conceptualization of this as a tool.

A few comments that just reinforce these comments. First of all, negative feedback in the pituitary is mediated by TR-beta 2. That needs to be specified there. We know this because -- well, I

have a bunch of references here that clarify that. The specific serum and clinical profile also of mutations in TR, are important. For example, even with TR-beta, a chemical that would activate TR-beta, you would expect to cause a reduction in serum TSH and T4.

But if you inhibit TR-beta, you would expect T4 levels to go up, but TSH, potentially, to remain the same. The complexity of the profile of blood levels of hormones for these kinds of interactions, needs to be, I think, carefully documented and incorporated into the analysis of the efficacy of these in vitro assays. Plus, I actually looked through the ToxCast database for these chemicals that hit, and there are a fair number of chemicals that hit both TR-alpha and TR-beta, sometimes with a low AC50.

Also, in this figure, the Delta T3 in cells and tissues needs to point to TR binding transactivation, not just gene expression. Because it works through changing thyroid hormone receptor binding. I'll leave it there.

DR. JAMES MCMANAMAN: Thank you. Dr. Belcher?

DR. SCOTT BELCHER: I have nothing else.

DR. JAMES MCMANAMAN: Dr. Furlow?

DR. J. DAVID FURLOW: I just wanted to amplify on a couple things. I agree with most of what was said. I think that this might seem it's not directly to the charge, but it does, so you have to give me a little…

The charge is asking us to assess how complete the AOP network is. Part of this is sort of asking us, is it sufficient to find the targets that the Agency may want to interrogate. At first glance, it looks complicated, and maybe overly complicated to someone looking at trying to link the estrogen receptor to an uterotrophic assay. But my argument about that is, is that I think we may be actually oversimplifying the estrogen and androgen pathways too much.

I think it's important for us to think about all the other ways that these steroid hormones could, in fact, be affected by chemicals. Including, of course, we are looking at steroidogenesis as well. I understand that. But you do have serum binding proteins, you have an alpha-fetoprotein, which is an

interesting challenge that may be bypassed or affected in some way. Feedback loops are even more complicated in the case of estrogen. Metabolism and clearance -- so this idea -- and I'll come to this in the second one about metabolism and clearance and how that may be affected too. And certainly windows of exposure or disruption during development, which is something I want to come back to, to the AOPs.

These are also, at least as important, I think, in the estrogen and androgen pathways. On one hand, one thing I was trying to -- it was an interesting pairing of looking more deeply at the androgen pathways to what we're being asked to do, to essentially reevaluate where the thyroid interrogation is at this point. It's interesting that I think the Agency -- I would urge the Agency to look at the lessons built on the estrogen and androgen programs. How that could be helpful, on one hand; but on the other hand, to use this AOP network and pathway that was, I think, elaborated well for the thyroid hormone system. I did want to say that upfront and I didn't, but I think it is elaborated well. But also essentially to investigate in the thyroid EDSP, or whatever.

The endocrine disruption of thyroid and hormone signaling in this way is a good context that could inform other ways that the Agency is operating. I think it's, you could say, do we need to look at all of these targets, do we need to look at all these AOPs, do we need to do all this stuff? I would argue on the face of it, yes, to a degree. And we'll talk about some of the MIEs in my charge. But I would also argue that it is instructive. I think it's a framework. I think it's a mental framework. It's a way to organize the way that these complex systems can be interrogated.

Anyway, so I think it came around to why one would want to look at that and how complete it is. I think it is fairly complete. There are other places that we could talk about, but I think it's very complete. I think it's important to be that complete.

The only other thing in terms of an AOP that might not be there is a challenge, again, for the steroid system as well, and that's to interrogate the effects and behavior. I've been on other SAPs where this has been brought up as well. I don't have a good assay for that, certainly not a high-throughput one. I think when we think about it, you could have

behavioral effects that are below even changes by histopathology in the brain. You could see subtle changes in branching of cerebellar neurons, these sorts of things.

I think it's important to think about that as an AOP, where subtle changes in thyroid hormone in those cells, at that time during development, is in fact the point. Inside those cells, what's T3 at that particular time? And whichever receptor that's causing those changes, and setting up that neural network.

Then that goes again to the critical period question, which I know the Agency does put upfront, but I think is missed sometimes when these AOP networks are set up. Those are just my two sort of general bigger picture comments essentially, towards the AOP question that was addressed to us. I'll just end there.

DR. JAMES MCMANAMAN: Thank you. Dr. Shaw?

DR. JOSEPH SHAW: I have nothing more to add other than to compliment the Agency on taking on this ambitious approach and just echoing David's

words of really trying to flesh it out in as much detail as possible.

Coming back to one minor point that Tom made, just keeping an eye on this broader literature that's already out there and coming out on thyroid hormone regulated transcription. Especially looking at newer tools that are coming out to really allay that into big networks of biological function.

DR. JAMES MCMANAMAN: Thank you. This charge question is now opened up for comments from other panel members. Yes, doctor?

DR. KRISTI PULLEN FEDINICK: I just had a question actually in terms of the critical periods. So, for thyroid -- and I don't know anything about this at all -- do the mechanisms or players change at all during development? Do you know that there is differential or different types of gene expression that's happening during development that wouldn't be captured in this current AOP network?

If this is just looking at adult biology, for the most part, are there things that we know are being up-regulated during development that would be impacted by, or impactful to, the thyroid signaling that aren't covered in this?

DR. THOMAS ZOELLER: Yes. It is true that it depends on what you mean by players, but certainly as thyroid hormone level declines during development, it has different effects on different brain areas at different times. In fact, the Agency -- - Mary Gilbert, in ORD, has probably published more kind of high-resolution information about that kind of issue than anyone.

In fact, I think she was probably the first person to really begin to look at the question about how low does thyroid hormone have to go before there is some change in the structure or function of the nervous system. I think that the Agency itself has really done a great job in telling us both how sensitive the developing brain is to low thyroid hormone, but also exactly your question. If you look in the hippocampus versus the cortex, you see different genes that are affected at the same temporal kind of period by low thyroid hormone.

It is complicated. I think it's going to be a real challenge and it's not the Agency's fault that somehow biology of the developing brain -- why would we think that that would be simple? It gets really complicated.

DR. KRISTI PULLEN FEDINICK: Just to follow up on that. Do you think that there would be different molecular initiating events, or is it really more downstream gene expression? I guess the question is, are there any pieces that are missing that are key events or MIEs that are unique to the developing body that you're not necessarily going to see in an adult animal or a fully developed animal? Maybe that's an unknown, which is fine. But I'm just curious, especially since development is very different than when you're not developing.

DR. THOMAS ZOELLER: I don't think we know that for sure. We don't have high resolution data. But we do know enough to propose that thyroid hormone -- the same thyroid hormone receptor in two different parts of the brain can have different effects. That's going to be mediated by differences in co-factor, differences in heterodimer formation, et cetera. There is some mechanism that controls that. I don't think we're ready for a key event that has that at the moment. It's speculative.

DR. J. DAVID FURLOW: I would just add that sort of, the development of competence for signaling is something that even was proposed way back

when, when they were putting neural tissues together and saying, okay, now we can respond.

The nature of that competence sometimes can be explained by receptor numbers. The tadpole tail has to have a certain number of TR-betas and then it does and it responds, and then the tail goes away, so, yes. That doesn't happen anymore. There are certainly some levels of thyroid hormone receptors in the adult frog, and yet we don't see much response.

It's going to be a combination, at times the receptor presence; it's going to be other times, as Tom brings up, even chromatin access to genes. So, the target genes could be the pioneering factors that open up genes and this sort of thing. It's, again, not the Agency's fault, it's our fault.

I think sometimes, yes, and sometimes, no, is the answer to your question unfortunately.

DR. JAMES MCMANAMAN: That was a discussion between Doctors Pullen Fedinick, Dr. Zoeller and Dr. Furlow. Other comments or questions?

Okay. Then back to the Agency.

DR. SEEMA SCHAPPELLE: I have no further questions for clarification, but I'll invite Dr. Lynn to ask any.

DR. SCOTT LYNN: First, I want to thank the panel for their comments. They are very much appreciated.

There are two things that I wanted to do. One, I wanted to ask, it sounded like the breadth of MIEs that were put forward were very much supported by the panel, but that there was a recommendation to better refine certain particular MIEs. In particular, the thyroid hormone receptor and the thyroid hormone transcription. I wanted to ask that as a clarifying question, is that what I heard?

DR. THOMAS ZOELLER: I think the point of that was to link the specific MIE. You can't just say, nuclear receptor or thyroid hormone receptor, because they're not uniformly linked to a particular adverse effect. The reason I'm saying that is that when you go to determine the balanced accuracy of some assay, you need to make sure that you're including, kind of, both sides of the equation. That is, of all the chemicals that affect serum T4 in guideline studies, how many of them work through this particular MIE? It may be that we would have to go into a higher resolution discussion about that calculation. The same will be true for the nuclear hormone receptors.

Even the NIS and TPO assay. That's kind of what I meant by that.

If you look at Table 4.1 and just generate one more column that has an adverse effect, or an effect, that is linked to the specific MIE -- you're going to have to separate TR-alpha 1, TR-beta 1, and TR-beta 2, because they do different things with respect to serum T4. The same is true for nuclear receptors, though, in the liver also.

DR. SCOTT LYNN: Thank you.

DR. JAMES MCMANAMAN: With that, I think we'll move on to Charge Question 6. We'll have it read into the record.

DR. RONNIE JOE BEVER: Charge Question 6. Please refer to White Paper Section 4.3. EPA has summarized currently available assays and test guidelines informative of thyroid adverse outcome pathways, and is developing high-throughput assays for a number of molecular initiating events.

Please comment on the ranked importance of molecular initiating events, Table 4.3, and whether assays for environmentally important molecular initiating events are missing; and include information

on both the biological and environmental relevance of these molecular initiating events.

In addition, please comment on other assays that would supplement, or be orthogonal to, the assays currently identified in Table 4.3, or for other key events or adverse outcomes in the thyroid adverse outcome pathway framework, Figure 4.2.

DR. JAMES MCMANAMAN: The panel members on this charge question were doctors Furlow, Belcher, Perkins, and Zoeller. Dr. Furlow is lead.

DR. J. DAVID FURLOW: As you might imagine, our discussions on this charge logically followed from discussions based on charge question five. The subgroup reviewed Section 4.3, discussed the proposed MIE targets for potential expanded screening, if necessary, and their ranking in terms of priority for the Agency, as well as if there were any potential missing MIEs that were worth discussing.

First, we found the coverage of the identified MIEs for the thyroid hormone endocrine system essentially comprehensive -- except for some of the granularity in terms of receptor subtypes that would be useful -- as outlined in Table 4.1. As I mentioned before, we appreciate the construct of AOP.

It's the best way to organize our thoughts to date. I think that's a very useful construct. The definitions used to describe the status of the assays, where they were essentially in a pipeline for each MIE, and the suitability for high-throughput, was fairly logical.

We would urge the Agency, however, to provide a clear definition of what high, medium, and low ranking means in terms of priority for action items and proposed timelines, for example. Does a medium ranking mean hold? We have assays we need for now? Or does it mean we have some assays and only need a few more orthogonal ones? Does low mean the assay would not develop assays until there is a hit from the literature, or an effect of a chemical on -- perhaps another example would be is that you wait until there is an effect on TH synthesis that's not explained by the existing TPO/NIS inhibition assays.

The other question that was raised, should priority be placed on MIEs that are most likely to cause a reduction in serum T4? Because that's where a lot of the chemicals that have been tested so far in Tier 1 and other essentially exploratory screening, has indicated. Is that the starting point? Do you start with the high-throughput assays that are

linked to reductions in serum T4? It could be best explained that way, was one other suggestion. What that ranking means -- I brought it up yesterday -- it still wasn't clear, I think, to us.

What I will do, though, is move on to commenting on each group of MIEs by the suggested priority by the Agency, including the suggestions then for the supplemental orthogonal assays where available.

In terms of the high priority MIEs, it makes sense, again, because the reduction in hormone is one of the things that's commonly observed that the sodium iodide symporter was relevant and thyroperoxidase were highly relevant.

In terms of alternative assays other than just measuring how radioactive iodide moves across a biological membrane, it's sort of hard to imagine other kinds of assays to come up with. I do agree in the general principle -- which came up in the first androgen receptor -- having different kinds of assays with different kinds of endpoints that are getting at the same biological question is important.

It was hard for me to think of one, other than to suggest potentially, that expression of

NIS in maybe xenopus oocytes as a model, which was sort of the old school idea. It is lower throughput, but it could provide more flexibility in examining different species' differences in NIS. There may also be polymorphisms in human NIS that could be investigated that way, and I think has been. This may be a faster way to explore NIS than creating new stable lines every time.

For thyroperoxidase, there are two assays currently available. The Tox21 data is available and I wanted to have a look, but didn't quite get a chance. I would say that I was happy the Agency did take care to examine -- when you have an assay that's based on loss of signal, you have to do quite a lot of controls. It seems like that the Agency is aware of that, but I would avoid those kinds of assays wherever possible. I couldn't think of a different way to do thyroperoxidase, that's not my area of expertise, but that was just my comment on thyroperoxidase. We felt those two merited being in the high category.

There was a special note and some discussion then about the hepatic nuclear receptor focus, and also the sulfation glucuronidation assays.

Where they fall in the high-minus and medium-plus range, it depends on how we define high and medium. In principle, the hepatic nuclear receptors would be important, A, because there is good evidence that when serum T4 is reduced, that liver metabolism is playing an important role, as was previously discussed.

Having a spectrum of assays looking at the hepatic nuclear receptors that are xenobiotic sensors could be important; but it sort of raises the interest, in some sense, if it is also of interest to the Agency to have a broad-spectrum way to interrogate those receptors. Because they could in fact be useful for looking at other lipophilic hormones as well.

I don't expect activating the hepatic nuclear receptors by one chemical or another is going produce -- doesn't produce just something that specifically metabolizes thyroid hormone per se. It kind of bumped up. We agreed that it is of interest as long as it has a broader attention, I suppose.

Species specificity is important there. We know that between rats and humans, that that can be very different, and so attention would have to be paid to that for sure.

I kind of lumped the hepatic nuclear receptor discussion with the sulfation glucuronidation assays under that same kind of category. It would be important to know, though, that when you are looking at -- so, these Phase 2 enzymes and the upregulation, potentially, of their activity, that you look at expression. That these are transcriptionally controlled by a lot of different hormones, and sometimes through hepatic nuclear receptors; and so not just the isolated enzymatic assays, but also expression.

It seems like, despite some of the challenges with the steroidogenesis assay that we discussed, that that's sort of an idea. That you would have a cell line that could look at both expression and activity of these enzymes all in one shot.

Some of the concerns that were raised about whether or not this is specific to thyroid hormone were addressed previously. At the end of the day, if thyroid hormones levels are reduced, this is an issue. It may be most relevant when feedback loops are not fully established during development, such as to the fetus. So, if you have thyroid hormone coming

internally, the animal in utero or the tadpole is starting to make its own thyroid hormone, but the negative feedback loop can't compensate quite yet, this may be where it's really relevant; and that's when a lot of these neural substrates are being affected.

The other one that was listed in the high category was the deiodinases. Certainly, the biology is clear, the genetic models support the basis for looking at them as important players for intracellular thyroid hormone, which at the end of the day is what is available to the receptors, and vary then closer to the biology, the downstream events.

The assays are currently in development, and so it was hard to then evaluate their suitability, their comprehensiveness. But certainly, the importance of the deiodinases to thyroid hormone physiology is essential. It is clear, and so it is potential that chemicals could affect them leaves it in the high category, and that makes some sense to us.

Moving into the medium category. Again, having issues with how things are ranked as high and medium. Kind of pulling together binding proteins, thyroid binding globulin, and transthyretin,

these are important modulators of free hormone, as well as distributing hormone to other systems. There are some chemicals that can dislodge, essentially, T4 from TBG, for example, so that could be important.

TTR is most important in amphibians for sure, versus mammals, so we should consider species differences there. The available assays -- it was raised, and we noticed this as well -- the high hit rate raised some concern. Additional validation of those assays, if they were to go into full consideration, should be paid attention to.

The transporters, in terms of the -- so I should have said distributor proteins. The membrane transporters, MCT8 for example, 10, the OATPs, LAT1 wasn't mentioned. That is another thyroid hormone transporter that is also an amino acid transporter. There is a clear genetic basis for looking at them and it's an emerging area of thyroid hormone physiology.

Certainly MCT8 has a very clear genetic link in humans. It's sort of in the range, I think, of the deiodinases, where the physiology is clear, the importance to -- you can have very differential expression of these transporters across the blood-brain barrier, for example. That sort of thing. It's

all important, but there really aren't assays yet. There's not really great candidates for how that might happen, although one could imagine they seem to be in development.

Whether or not looking at the TRH receptor, TSH receptor, again, seemed reasonable, was placed in the medium category. Species considerations are important here, as was raised for some of the works on amphibians. So, corticotrophin releasing hormone does drive metamorphosis in amphibians. This has been examined, and so CRH receptor could be another point of consideration. If the Agency is looking at the corticoid signaling pathway at some point, that that may be incorporated there.

I have most to say about thyroid hormone regulated transcription. The considerations that thyroid hormone regulated transcription falls into the medium category makes some sense as a priority on one hand, because we do have reporter assays that have been used in high-throughput. There are some candidates out there. The conclusion is that the receptor is rather finicky about what it binds to, and that's fine. That makes sense to us as well.

We did already discuss, though, about the issue of really examining TR-alpha and TR-beta. The existing assays for TR-alpha are mostly overexpression and GAL4 based. One thing to keep in mind, is that the idea is that there is not -- there are some synthetic compounds. There is some evidence here and there that there may be some differences in the way chemicals interact with the TR-alpha and TR-beta binding pocket, but not a lot. But it is certainly true that within chromatin, within specific target cells, the proper assembly of complexes may differ and modulate ligand potency via different isoforms.

Looking at TR-alpha versus TR-beta and the pathways that they regulate, the transcriptional networks that they regulate would not necessarily be revealed by overexpress receptors, and certainly GAL4 fusions, which is usually what is present. One thing we considered, in thinking about the transcriptional assays and if they were sufficient or if additional work needed to be done, is that you can move either in one of two directions.

One, you can look at specific downstream target genes, if in fact, they are

identified as a key event. So, if there is some evidence that a common downstream gene that is tightly regulated by thyroid hormone is a relevant target gene that then was linked to the AOP. KLF9 is suggested as one potential to look at. I think now that the technology is moving, you don't want to be tied necessarily to KLF9 if it's upregulated, say, in the liver, but doesn't do very much. Although evidence suggests it does.

If you can now develop newer, high-throughput transcriptomics, RNA-seq, that sort of thing, that may be something to consider and a way to move.

One thing I wrote here is that if I knew then what I know now in developing reporter gene assays; or could do now as opposed to could do then -- both financially as well as technology -- this would be something we would recommend. Reporter genes essentially were at the birth of the nuclear receptor field. They have been very useful for a number of years. A lot of the Tox21 assays are built on these reporter genes, but they are pretty artificial themselves.

I just gave a little brief about GAL4 fusions. Well, minimal promoters and luciferase assays or beta lactamase assays have other issues as well. I think interrogating -- we think. There were multiple suggestions from the group to look at incorporation of targeted high-throughput RNA sequencing in amenable cells -- and that's going to be another trick -- or trackable organisms, for the identification of activated pathways relative to thyroid function and disruption would be recommended as orthogonal or even replacement assays for the old reporter gene assays.

Where to do this? You could do this in GH3 cells. They are responsive. That's where the luciferase reporters were built. But such techniques don't tie you to making a stable reporter line. Now you can screen primary cell lines, IPSC derived specific cell types, but only if they retain appropriate thyroid hormone responsiveness, which can be a trick.

Liver, but certainly neural models might be the most important places to start, but thyroid hormone is important for many tissues and organs.

Along these lines, animal models might be useful if adopted for this kind of approach. Xenopus laevis can be miniaturized to at least medium-throughput, perhaps, since they are competent to respond very early to exogenous hormone. The genome is now complete, so high-throughput transcriptomic approaches are feasible here. Zebra fish also has potential for medium-throughput animal model. I would argue that the thyroid endocrine physiology is less fully understood in zebra fish, but that provides a potential as well.

There is also an opportunity, moving along these lines, to look at cells and look at animal models to use genome editing in these simpler, less expensive models, to develop hypothesis driven testing of chemical affects. This could help ID key events and fill in gaps between MIEs, in particular, AOPs, which is a goal of what the program needs to do here.

Any use of cells or animals with phenotypic or transcriptional readouts, will need to have fully characterized transporters, deiodinase activities, both temporally, spatially, that will all have to be characterized, and I will make a point about that later on.

Our next move to the lower priority MIEs is a summary. Receptor binding itself in isolation was agreed to be less informative. Affinity sometimes is not representative in purified proteins overexpressed in bacteria or something like this. If the confounds can be removed, and it can be made more convenient, we think that the transcription readouts are superior.

It was suggested that other steps of TH synthesis, such as pendrin, dual oxidase, iodotyrosine deiodinase, ranked lower for the time being. We agreed that if the assays are not there, if there is not evidence for a chemical affecting them at the moment, that looking at NIS and TPO to see if that covers a broad spectrum of chemicals of concern affecting thyroid hormone levels in vivo, is probably a way to go. But paying attention to the literature is probably a good idea.

In terms of the question of missing assays. The committee discussed the following. For any assay that has been brought up, biotransformation, I think, is important. So, linking what you're doing to chemicals to see if they are, first, hydroxylated. There are certain key hydroxylation steps in flame

retardants that allow the flame retardant to even bind to the receptor. So, that's one step that also is linked to subsequent sulfation and glucuronidation. Not binding to the receptor, but actually setting up those reactions. Somehow linking that to biotransformation, I think, is important.

The other one that is emerging, I think, is retinoid X receptor ligands, both pharmaceutical and environmental. It's been known for a while that synthetic pharmaceutical RXR ligands suppress TSH and make patients hypothyroid. So, they were originally promising for certain cancers, but then were stopped because they were causing severe hypothyroidism. So, RXR itself -- whether or not that's in combination with TR unknown -- but certainly RXR could have effect on thyroid hormone. It also has effects on metamorphosis; ligands for RXR have effects on metamorphosis as well.

One other target or one other concern that was brought up in thinking about different kinds of assays or things we need to be concerned -- we thought the Agency should be concerned about, is that -- I just learned this this week -- that lithium -- still used to treat bipolar disorder -- leads to

hypothyroidism in a significant number of patients. But the mode of action is not 100 percent clear, but there is evidence that it is actually linked to thyroid hormone release. So, lysosomal coupling and release and so what is lithium doing in there? There are some suggestions, but that's another potential MIE to keep in mind.

Overall, linking the MIEs to key events, to AOPs, and looking at the quantitative differences through these pathways using cell lines or animals to assist this goal, we felt was important. But linked to an earlier comment I had, I think if there is a way to hit multiple MIEs at once, if there is a cell line that expressed deiodinases, also has a certain array of transporters and has a nice transcription readout, sort of the unicorn of thyroid hormone signaling, that would be awfully nice, but they may exist. And they may not be a cell line, it may not be an IPSC, but it may be a small organism -- model organisms.

It was also suggested that other species are missed by focusing on essentially the lab rat of the amphibian world, which is xenopus, or zebra fish, the lab rat of the fish. That we should

consider, potentially, other organisms. There is some thought that thyroid hormones or iodotyrosine may play a role in sea urchin metamorphosis. These are also organisms of concern in the marine environments, that sort of thing. Other species should be potentially considered for screens, but at least consideration by the Agency at some point.

Those are most of my comments. I think that's pretty much where it is. There were other suggestions about how critical it is, as the Agency is aware, to develop reference chemicals along with selected MIEs. As you're developing the assays, to have some good reference chemicals. It's noted that a specific TR antagonist is now more widely available than it had been. That may be useful.

Species differences are implied in the charge question and the environmental relevance. And we discussed a little bit about that, but maybe this isn't quite where we can approach this. But genetic variation and sensitivity to hormones and chemicals and how it may affect the thyroid hormone endocrine system is not captured this way, and where would this reside? Where can this be interrogated and when? Is it all pharmacokinetics? Is it target cell

sensitivity? Humans typically have a tighter range of T4 than circulating T4 levels. So, the points can be different and not understood. But that may be beyond the scope of what we're asked to do here.

I'll stop there and turn it over to my colleagues.

DR. JAMES MCMANAMAN: Thank you. Dr. Belcher?

DR. SCOTT BELCHER: That largely covers everything that I had thought about, although as this went on, I just wanted to make a comment on the adverse outcomes that we're potentially missing. I think the Agency has a real opportunity here to address a long-standing hole in assays that relate to neurodevelopmental effects. This may be a real opportunity here to focus on filling that void.

DR. JAMES MCMANAMAN: Thank you. Dr. Perkins?

DR. EDWARD PERKINS: No, they've all covered everything that I was concerned about.

DR. JAMES MCMANAMAN: Dr. Zoeller?

DR. THOMAS ZOELLER: Yes. Two things. One is in terms of adverse effects, I think myelination and oligodendrocyte development is

actually both important in vivo, but also in vitro. There are some in vitro assays that I think could be really useful there.

The second thing is this idea of prioritization, because as I think about what high-throughput assays related to thyroid are going to do for the Agency, and if I compare that to estrogen, androgen, steroidogenesis, it's what's available in EDSP Tier 1 that could be captured here. Really, essentially, what you're looking to replace is T4 assays in the pubertal assay.

That's really the first thing to focus on in terms of being practical, because that's going to have some kind of downstream effect that's going to have some utility for what you're doing. Some of the MIEs that you identify are related to serum T4, some aren't. I think that to be practical is kind of useful here.

I also think that it's clear that there are many chemicals -- well, I can't really say many. I can't quantify that. But there are chemicals that can affect thyroid hormone action in tissue without affecting thyroid hormone levels in blood. That's a

deeper problem that is going to require a lot more effort.

DR. JAMES MCMANAMAN: Thank you. This charge question is now open for comments from other panelists. Hearing none, I'll go back -- why not?

DR. KRISTI PULLEN FEDINICK: I'll just ask one. I had one written down. I wasn't going to ask it because I wanted to shake things up a bit.

I guess the question was about -- and we talked a little bit about -- this is Krisi Pullen Fedinick from NRDC. We talked a little bit about this yesterday and I think it was just brought up in Dr. Furlow's comments as well, about the utility of the high, medium, and low. I think that there is some practicality to that to get to Dr. Zoeller's comment.

But then I also just wonder about the need for doing that -- maybe it just comes down to explanation of why it is that these are being prioritized in this way. But if you said for pendrin, and dual oxidase, and iodotyrosine deiodinase, those are low for the time being, but they should be included later.

Were there any of the other ones in here that you would say -- because the low were being

thrown out of the second figure with the AOPs, right? So, that was one definition that we had essentially, that these would not be looked at right now, and maybe not even in the future. Were there any that you would say that that should not be the case? Does that include those three in the biosynthesis in the thyroid? Does that include that alanine side-chain?

If we were going to throw out MIEs, are there any that you would say are okay being thrown out for the long-term? Or should all of these be included in the long-term? I guess that's the clarifying question.

DR. J. DAVID FURLOW: Well, from my perspective as a basic biologist, I'm kind of interested in all of them, right? And so both genetic models and pharmacological ways to manipulate their activity, and so I have sort of have a larger interest rather than the screening purpose.

Are they equal in their biological significance? They're all part of the pathway. So, pendrin has other effects, effects on hearing, for example. So, there is a syndrome that causes deafness, and why would -- I mean, it's an anion transporter.

Some of them are categorized, I think, would be in my mind a little bit lower because they are broader. They have roles outside of thyroid hormone -- so basically you just have to get the iodide right across the basolateral membrane and get it into the colloid. It does a good job, and you would be hypothyroid if you didn't have pendrin, but you'd also be deaf -- actually, that could be independent from being hypothyroid.

LAT1 is another transporter. MCT8 is up there. Some of them are general amino acid transporters that if you -- the knockouts, for example, are not very healthy, so do you knock it out specific -- you know. That's where I'm sitting on that. Should they never be looked at ever? I guess that's a stretch. I guess I wouldn't suggest the Agency looks at low priority that way.

I think there should be -- I guess one thing to suggest -- and maybe I didn't frame it quite the right way about high, medium, and low ranking -- was about, can you place that into a decision tree? I think that was raised by one of the public commenters. I don't know that that was right, you know? What does it look like to be high, medium, and low in terms of

decision tree for testing? I think that's an important point. I don't have the answer, but I guess the recommendation would be that the Agency would be clearer on that. Would I throw anything away? Well, no.

DR. JAMES MCMANAMAN: I would like to see his garage. Dr. Zoeller?

DR. THOMAS ZOELLER: I'm going to respond to this also because I think it really goes to this issue of prioritization. I agree. I wouldn't throw anything away either, but when you look at the full constellation of data, for example, look at MCT8 versus TPO. Humans that have an MCT8 defect have severe neurological deficits that cannot be overcome. Humans with the TPO deficit, at birth, are identified in a screening program and given thyroid hormone and they fall within a normal range of functionality.

In terms of mice, if you knock out MCT8, well, they don't really have a phenotype because they also have other transporters that can compensate for that loss. From the Agency's point of view, if you're going to use an assay for MCT8 and compare that to rodent studies, I think it's going to be

complicated. It's going to be asymmetric, so it's not going to match.

The Agency has to really look at those kinds of details -- I'm sure they do -- to prioritize what they're going to do. I also wouldn't throw anything out because all of these MIEs are important for the regulation of thyroid hormone action. I actually hate the word homeostasis because it doesn't really mean anything. I think that the Agency needs to be practical about what can be done sooner rather than later, and what the purpose of that is going to be.

DR. JAMES MCMANAMAN: Thank you. Dr. Clewell?

DR. REBECCA CLEWELL: Why not, right? I want to make a bit of a point here that maybe piggybacks a little bit off of what Dr. Furlow was talking about in terms of the complexity right now of the proposed AOP for thyroid, versus the overly simplistic view of the estrogen or androgen pathways.

I feel like that was important because actually probably what we need to do is titrate to get somewhere in between that to a happy medium. If I wanted to, give me two hours, and I'll make a diagram

like that for estrogen or androgen where there are just as many nodes. It's not saying that -- I know it was hard work and you did that over years, so I'm not trying to minimize that in any way. But for several years I've been mapping the estrogen pathway. I do have an AOP that looks very, very complex for estrogen. I think I'm publishing it. Anyway, but I may have already, I can't remember. I think I did in a review article.

The point is, though, our goal, I believe, as scientists moving towards an in vitro approach, is to say, use the principle of parsimony. How much do we need to include in order to get what is most important for the risk assessment decision? So, in the case of estrogen -- while I would personally like it to be more complex, and there are things that are not currently being considered right now that I would like to have considered, we do have most of the sort of key determinants of whether or not we're going to see a phenotypic response for the more common endpoints.

But, we haven't begun to address the fetal situation. We haven't begun to address the tissue specific situation with estrogens. Nobody

talks about bone. I don't know. I like the idea of having a tiered approach that the EPA has laid out here; where we have a prioritization of really high priority targets because we believe these are quickly going to get us to a screening approach for the kind of larger issue. Then we can move into the more specific responses, the more specific MIEs if we need to.

The kind of one thing I want to tag onto that, is if we look at what we have for estrogen, we have -- with whatever flaws may be there -- we have the nuclear receptor response and transactivation. We have steroidogenesis. We never talk about the fact that metabolism is just as important for any of the others as it is. I mean, estrogen is cleared through sulfation and glucuronidation. And certainly, we can disrupt estrogen hormone that way. We need to add that, but we have the first two.

We could have that same thing with thyroid and it would get us like 80 percent of the way there. If we had thyroid hormone synthesis, thyroid hormone metabolism, we would be just as far along as we are with estrogen. Unfortunately -- and I've just Googled it again. I Google it about every two months

to see if there is a thyroid hormone synthesis assay. So, there is not. I guess I'm pulling the same thing as Dr. Pennell did. There is a problem and I don't have the answer for it.

Wouldn't that get us a lot of the way there if we have five to 10 molecular initiating events that lead to a reduced thyroid hormone synthesis? What if we could measure thyroid hormone synthesis in a phenotypic assay? It would really reduce the anxiety about whether I had a pendrin assay and I had a TPO assay and I had a whatever assay.

That's my two cents on that. Not to minimize the complexity of the thyroid, because it is really complex. I was super impressed when I saw that AOP. Because I was like, man, a lot of work went into that. I wish more AOP networks like that were built. I just don't want to also get to the point where we have such complexity laid out on a scheme that we start thinking, man, we're going to have to have 342 assays just for this one network is all.

DR. JAMES MCMANAMAN: Thank you. Other comments? Hearing none, I'll go back to the Agency.

DR. SEEMA SCHAPPELLE: I just want to thank the group and the panel for comments. Not just

on this section, but all day. And from yesterday as well. I think that really underscores our history of innovation within the EDSP and that's not intended to stop. I think this information that we've gotten really helps us with that, and doing that with the purpose for meeting the mission of the Agency, protecting human health and the environment. So, thank you.

DR. SCOTT LYNN: I want to reiterate what Dr. Schappelle just said. And I also, for this last charge question, I want to thank all the panel members for their input. It was very illuminating.

There are a few things that I do want to say. Number one, that AOP diagram was actually made by probably like 14 scientists at the EPA. It will be a publication that is coming out soon. I don't want to take credit for it. It wouldn't be appropriate for me to take credit for it.

There are a few other things that I wanted to make sure that I correctly heard. The ranking prioritization on MIEs, I heard positive feedback for that, but there was sort of a suggestion that it's more clear, or there is a decision tree that would be associated with that. Is that correct?

DR. THOMAS ZOELLER: Yes, that's what I was thinking, yes.

DR. SCOTT LYNN: Thank you. Then I would ask in the minutes, that are put together for this charge question, that you expound upon some of the things. The LAT1, the KLF9, getting that input will be highly valuable. I thank you for that.

Also, addressing RXR ligands and their role, I think would also be very important to get that into the panel's response.

I also heard a recommendation for neurodevelopmental assays and pursuing that. I wanted to thank you for that.

Is there anything else?

DR. REBECCA CLEWELL: Thyroid hormone synthesis.

DR. SCOTT LYNN: Yes. Thank you. I'm not going to stop this meeting from ending. Thank you very much.

DR. JAMES MCMANAMAN: I guess that completes the charge questions. At this stage, what we typically do is go around and if there are any last comments. I think I'll start.

I want to really thank the Agency for some very clear presentations on some very complex materials and for giving me new insight into how to do some kinds of assays that I think this -- I can't say the word yet -- the Mahalanobis assay. I think it has a lot of equitability and I know the statisticians and mathematicians, or whatever they'd like to call themselves, think that it's a general assay, but I have never heard of this before. I think it's very cool. I thank the Agency for that.

I think you guys are really on the right track in developing your high-throughput assays, and think that you're very close to -- at least from my perspective -- you're very close to having those very workable. It seems like you made a lot of progress there since the last meeting that we had.

With that I'll thank you again. It was a very nice presentation and very informative.

Dr. Barr?

DR. DANA BARR: Thank you. It was great presentations. I learned a lot. Sorry I didn't have a whole lot to add to it as this is not my real area of expertise. But very great job doing this, and

it's good to see that you're moving forward with alternative assays.

DR. MARION EHRICH: I thought your document was well written and actually quite easy to read, which I appreciated very much. And the presentations, of course, were good. I have to give you compliments for attacking that thyroid problem, because that's going to be a really hard one. That you actually made a step in that direction is appreciated.

DR. JOSEPH SHAW: It's starting to sound like a cabinet meeting here. I just want to second what everyone said, and thank you again for the completeness of what you've put together.

DR. SONYA SOBRIAN: Hi. I'd like to thank you for all of your hard work and to acknowledge the interagency cooperation that the document reflects. And as usual, I've learned a lot. Again, this is not really an area that I work in and I've really learned a lot. Thank you very much.

DR. SUSAN NAGEL: Yes. Ditto. Seriously, I appreciate the huge amount of time and commitment that you all put into this. I know it's not easy. I guess my one request would be going back

to looking at the sensitivity of some of your assays. I would just encourage a little bit more rigorous comparison of sensitivity, and perhaps a little bit more transparency around that. Thank you.

DR. THOMAS ZOELLER: I agree with everybody.

DR. GRANT WELLER: I would just like to also thank the Agency for the clear presentations. I guess I learned a lot in addition to maybe Dr. McManaman knowing the difference between a mathematician and a statistician. I think some of the work presented here today -- without knowing the biology behind it, it's a really cool example of using new technology, and data, and smart data science to create a lot of value and be able to do things more efficiently and using lower resources. That's really impressive.

DR. KRISTI PULLEN FEDINICK: I really appreciate you putting up with my questions and comments and things like that. I really think that the Agency is on the right track and you have some really beautiful and powerful tools at your disposal. I appreciate the opportunity to be able to offer insight and input into those, and so making this a

public process I think is really helpful. Just continuing to do good work and ensuring that the tools are ultimately protective of public health. I just appreciate this opportunity.

DR. EDWARD PERKINS: I just reiterate what everyone else says. I think you did a great job. The documentation is very helpful, especially the introduction of the accessing documents through HERO. That really facilitated in looking at some of the more obscured documents to kind of track down and make sense of some of the things that were going on. Thank you very much.

DR. REBECCA CLEWELL: I should also say thank you because I've had a lot of comments, and you guys take them all very patiently. I think what I would like to say is that I'm 100 percent in support of moving these assays forward. And while I might have some tough questions, it's because I'm thinking critically about it. I want to make sure they are very strongly supported, because I'd like to see them used more broadly in the community and ultimately towards a risk assessment purpose.

I think the steps that are being made here, what has been done with estrogen, and what is

beginning to happen with steroidogenesis and even thyroid and androgen, is -- every move forward is a good move forward for us. As long as it is in the public health interest. I do believe that the assays that we are discussing today are.

I also appreciate all of the work that went into that, because that document was enormous and very well documented. I cannot believe the amount of documentation that we were given, so I know that that was a tremendous effort. Thank you all.

DR. MICHAEL PENNELL: I'd like to commend the EPA for their hard work and I'd also like to thank them for developing new technologies which presents new statistical problems and keeps us busy and employed. Thanks.

DR. IOANNIS ANDROULAKIS: For me, this was the first time that I had the opportunity to participate in a panel like that, I have to say it was like a great and very impressive experience. So, thank you.

DR. SCOTT BELCHER: I would pretty much say, ditto, to everything else that has been said. But I'd also like to commend the Agency on its mobility and its embracing new technologies and

approaches for toxicological testing. This has been one of the -- getting stuck in previous technology and assays that have been validated, et cetera, et cetera, and the slowing pace that does not embrace current science. I appreciate these efforts.

DR. VERONICA BERROCAL: I wanted to thank you, first of all, for giving me the opportunity to be on this panel. I always appreciate coming to these panels and see actually public health in action as I like to think about it. I also really want to commend everybody for the great work. The creativity that you guys have is just amazing and you do science that encompasses all different disciplines, so that's really great to see.

DR. J. DAVID FURLOW: I want to spend about 15 minutes on these thyroid organoids if we can -- maybe after, right. A beer may be good.

I do want to thank the Agency as well. Just one thing that maybe you've heard, and maybe you haven't. To me this particular program -- it may be true of all the Agency science I hope -- but it's an opportunity to have basic scientists, academicians, interacting with folks that do very applied and very

sort of -- very public-facing work. To me, what's exciting about it is that it informs my basic science.

There are ideas and challenges that are put forth by the Agency to develop assays and to think about prioritization, and what's important and what's not important, what do the genetics say, what do the -- I think that's a -- maybe it's all obvious to all of you. But to me it's something that I've really appreciated over the years, is being able to interface in that way because it's helped my basic science and my excitement about what I do and why I do things. I just wanted to point that out. Thanks.

DR. TODD PETERSON: I would just like to acknowledge that as a FACA event, we have members of the public who have been on the phone and in the room and present, and I'd like to thank them for listening in. I also want to thank the commenters for their contributions. It appears that they have been heard, and supplemental materials that have been brought to our attention during the meeting will certainly go into the docket.

In the process of assembling the expertise and the diversity of disciplines and whatnot, it's been interesting for me. I think that

we have had the appropriate mix of ad hoc and permanent panel members to contribute to a good a robust dialogue and a collaborative conversation. I thank you for coming far and near to be here for this.

I may have forgotten something else. I apologize for that. But if nothing else, if the chair -- which I'd like to thank Dr. McManaman for being the chair -- if he concurs, then we can draw this meeting to a close.

DR. JAMES MCMANAMAN: I concur. I would like to have all the panelists in our breakroom for our short meeting, post-meeting wrap-up. Thanks, everyone.

[WHEREAS THE MEETING WAS ADJOURNED]

* * * * *